Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

МОДЕЛЮВАННЯ МЕТОДУ ПРАВДОПОДІБНОЇ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ПОТОКІВ ДАНИХ В ОНЛАЙН РЕЖИМІ
(тема)

Виконав:
здобувач 4 року навчання,
групи ІТІНФ-21-1
Ніколаєва Д.Ю.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Шафроненко А.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Завідувач кафедри інформатики _____ Кобилін О. А.
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту

Кафедра Інформатики

Рівень вищої освіти перший (бакалаврський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Ніколаєвій Дар'ї Юріївні
(прізвище, ім'я, по батькові)

1. Тема роботи Моделювання методу правдоподібної нечіткої кластеризації потоків даних в онлайн режимі

затверджена наказом університету від 19 травня 2025 року № 381Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 09 червня 2025 р.
3. Вихідні дані до роботи науково-методична та науково-технічна література з тематики нечіткої кластеризації, статистичної обробки даних і потокового аналізу; матеріали наукових конференцій з інтелектуального аналізу даних; електронні ресурси інтернет-мережі; публікації з відкритим доступом щодо алгоритмів онлайн кластеризації; синтетичні та відкриті реальні набори даних; інструменти математичного моделювання та бібліотеки для машинного навчання.

4. Перелік питань, що потрібно опрацювати в роботі
1. Обґрунтування актуальності задачі кластеризації потоків даних в онлайн режимі.
2. Огляд методів нечіткої кластеризації та їх адаптацій для потокових даних.
3. Формалізація методу правдоподібної нечіткої кластеризації та математична модель.
4. Програмна реалізація методу та організація обчислювального процесу.
5. Експериментальне дослідження та порівняння ефективності з існуючими методами.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)
Правдоподібна нечітка кластеризація потоків даних, особливості потоків даних, мета та завдання дослідження, теоретичні основи нечіткої кластеризації, математичне моделювання методу, програмна реалізація та експерименти, порівняльний аналіз методів, висновки та перспективи.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	07.04.2025	
2	Аналіз завдання, підбір літератури	08.04.25-10.04.25	
3	Аналіз літератури з досліджуваної проблеми	11.04.25-14.04.25	
4	Аналіз існуючих методів	15.04.25-20.04.25	
5	Розробка методу	21.04.25-27.04.25	
6	Програмна реалізація	28.04.25-11.05.25	
7	Оформлення пояснювальної записки	12.05.25-20.05.25	
8	Перевірка на нормоконтроль	21.05.25-01.06.25	
9	Перевірка на плагіат	21.05.25-01.06.25	
10	Рецензування	21.05.25-01.06.25	
11	Підготовка презентації та доповіді	21.05.25-18.06.25	
12	Занесення роботи в електронний архів	02.06.25-18.06.25	
13	Попередній захист кваліфікаційної роботи	02.06.25-18.06.25	

Дата видачі завдання 7 квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Шафроненко А.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 75 с., 9 табл.,43 джерела.

НЕЧІТКА КЛАСТЕРИЗАЦІЯ, ПОТОКИ ДАНИХ, ПРАВДОПОДІБНІСТЬ, ФУНКЦІЯ НАЛЕЖНОСТІ, ІНКРЕМЕНТАЛЬНЕ НАВЧАННЯ, ОНЛАЙН РЕЖИМ, ГАУССІВСЬКА МОДЕЛЬ, АДАПТИВНІ МЕТОДИ.

Об’єктом роботи є процеси кластеризації потоків даних в онлайн режимі з використанням методів нечіткої логіки та статистичної теорії оцінювання.

Метою роботи є розробка та моделювання методу правдоподібної нечіткої кластеризації потоків даних, що забезпечує ефективний аналіз неперервних послідовностей спостережень в онлайн режимі з урахуванням невизначеності та статистичної природи даних.

У результаті роботи здійснена програмна реалізація методу правдоподібної нечіткої кластеризації потоків даних та його експериментальна верифікація на синтетичних і реальних наборах даних.

FUZZY CLUSTERING, DATA STREAMS, LIKELIHOOD, MEMBERSHIP FUNCTION, INCREMENTAL LEARNING, ONLINE MODE, GAUSSIAN MODEL, ADAPTIVE METHODS.

The object of the work is the processes of clustering data streams in online mode using fuzzy logic methods and statistical estimation theory.

The aim of the work is to develop and model a likelihood-based fuzzy clustering method for data streams that provides effective analysis of continuous observation sequences in online mode, taking into account uncertainty and the statistical nature of the data.

As a result of the work, software implementation of the likelihood-based fuzzy clustering method for data streams and its experimental verification on synthetic and real datasets were carried out.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

FCM – Fuzzy *C*-Means (метод нечіткої кластеризації середніх)

PFCM – Possibilistic Fuzzy *C*-Means (можнісна нечітка кластеризація)

GMM – Gaussian Mixture Model (модель суміші гауссівських розподілів)

EM – Expectation-Maximization (алгоритм очікування-максимізації)

ARI – Adjusted Rand Index (скоригований індекс Ранда)

NMI – Normalized Mutual Information (нормалізована взаємна інформація)

AIC – Akaike Information Criterion (інформаційний критерій Акаіке)

BIC – Bayesian Information Criterion (інформаційний критерій Байєса)

PC – Partition Coefficient (коефіцієнт розбиття)

CE – Classification Entropy (ентропія класифікації)

XB – Xie-Beni Index (індекс якості кластеризації)

τ – порогове значення для виявлення аномалій

$\mu_i(x_t)$ – ступінь належності t-го спостереження до i-го кластера

x_t – t-те спостереження потоку даних

v_i – центр i-го кластера

Σ_i – коваріаційна матриця i-го кластера

π_i – апріорна ймовірність належності до i-го кластера

z_{ij} – бінарна змінна приналежності x_i до кластера j

$\ell(\theta|X)$ – логарифмічна функція правдоподібності

λ – параметр забування ($0 < \lambda < 1$)

Δ – розмір часового вікна

$W(t)$ – множина спостережень у часовому вікні

Concept Drift – дрейф концепцій (зміна структури даних у часі)

Single-pass – одноразове сканування даних (без збереження історії)

OnlineEM – онлайн-алгоритм очікування-максимізації

StreamReader – модуль зчитування потоку даних

FCMUpdater – модуль оновлення параметрів кластерів

t-розподіл – статистичний розподіл з важкими хвостами

ВСТУП

Стрімкий розвиток інформаційних технологій та цифрова трансформація усіх сфер людської діяльності призвели до експоненційного зростання обсягів даних, що генеруються та обробляються в онлайн режимі. Потоки даних, що характеризуються безперервністю, динамічною природою та потенційно необмеженим обсягом, потребують розробки спеціалізованих методів аналізу, здатних обробляти інформацію в реальному часі без збереження всієї історії спостережень [1, 2].

Кластеризація потоків даних є одним із фундаментальних підходів до виявлення закономірностей у неперервних послідовностях спостережень. Традиційні методи кластерного аналізу, розроблені для статичних наборів даних, виявляються малоефективними при роботі з потоками через їхню обчислювальну складність та нездатність адаптуватися до змін структури даних у часі. Це створює необхідність розробки спеціалізованих методів кластеризації, орієнтованих на роботу з потоками даних в онлайн режимі.

Нечітка кластеризація, що дозволяє об’єктам належати до кількох кластерів одночасно з різними ступенями належності, надає гнучкий апарат для моделювання складних структур у даних та врахування невизначеності. Однак, класичні методи нечіткої кластеризації потребують багаторазового проходу по даних та збереження всієї інформації в пам’яті, що неможливо в контексті потоків даних. Це зумовлює необхідність адаптації принципів нечіткої кластеризації для роботи в онлайн режимі.

Правдоподібна кластеризація, що базується на статистичній теорії оцінювання та використовує функцію правдоподібності для формування кластерів, надає формальну математичну основу для кластерного аналізу та забезпечує статистичну інтерпретацію результатів. Інтеграція принципів правдоподібності в нечітку кластеризацію потоків даних дозволяє поєднати переваги статистичного підходу та гнучкість нечіткої логіки для ефективного аналізу неперервних послідовностей спостережень.

Моделювання та програмна реалізація методу правдоподібної нечіткої кластеризації потоків даних в онлайн режимі дозволяє оцінити ефективність запропонованого підходу в різних умовах та для різних типів даних, а також порівняти його з існуючими методами.

1 ТЕОРЕТИЧНІ ОСНОВИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ПОТОКІВ ДАНИХ

1.1 Поняття та особливості потоків даних в онлайн режимі

Потоки даних являють собою безперервні послідовності інформаційних елементів, що надходять із зовнішніх джерел та обробляються системою в режимі реального часу. Потік даних характеризується динамічною структурою та необмеженим обсягом, що унеможливлює збереження всієї інформації протягом тривалого періоду. Обробка потоків даних в онлайн режимі передбачає аналіз та реакцію на інформацію без зупинки процесу отримання нових елементів даних, що робить цей підхід незамінним для систем реального часу.

В контексті аналізу потоків даних в онлайн режимі застосовується принцип одноразового сканування (single-pass scanning), коли кожен елемент даних опрацьовується системою лише один раз без можливості повторного доступу до нього [3]. Така особливість зумовлена постійним надходженням нових даних та обмеженнями щодо обсягів пам’яті, що можуть бути виділені для зберігання історичних даних. Онлайн обробка потоків даних спирається на адаптивні алгоритми, здатні пристосовуватись до змін структури та характеристик даних без перезапуску процесу обробки.

Концептуальна модель потоку даних визначається як нескінченна послідовність елементів $X = \{x_1, x_2, ..., x_n, ...\}$, де кожен елемент x_t належить багатовимірному простору ознак. Потоки даних мають часовий аспект, оскільки кожен елемент асоціюється з моментом часу його надходження, що створює темпоральну залежність між елементами та може впливати на процеси аналізу даних [4, 5]. Часова прив’язка елементів потоку створює підґрунтя для виявлення закономірностей еволюції даних у часі.

Основними характеристиками потоків даних є: високий обсяг надходження даних, висока швидкість їх обробки, ймовірна зміна структури даних, наявність шумів та аномалій, нестаціонарність потоку [6, 7].

Нестаціонарність проявляється через зміну статистичних характеристик даних з плином часу, що призводить до явища дрейфу концепцій (concept drift) – змін закономірностей та взаємозв’язків між атрибутами даних. Виявлення та адаптація до дрейфу концепцій становить одну з суттєвих проблем аналізу потоків даних в онлайн режимі.

Архітектура систем обробки потоків даних розподіляється на декілька рівнів: рівень збору даних, рівень попередньої обробки, рівень аналізу та рівень прийняття рішень [8]. Рівень збору даних забезпечує отримання інформації з гетерогенних джерел та формування єдиного потоку даних [9]. Рівень попередньої обробки виконує фільтрацію, нормалізацію та інші трансформації даних для підготовки їх до аналізу. Рівень аналізу реалізує алгоритми видобутку знань з потоку даних, а рівень прийняття рішень використовує отримані знання для генерації керуючих впливів.

Специфіка роботи з потоками даних в онлайн режимі обумовлена необхідністю обробки даних із дотриманням часових обмежень. Обчислювальна складність алгоритмів має бути пропорційна обсягу даних, що надходять за одиницю часу, щоб забезпечити своєчасність реакції системи на вхідну інформацію. Обмеження на використання обчислювальних ресурсів зумовлює застосування апроксимаційних методів та евристик, що дозволяють знайти компроміс між точністю та швидкістю обробки даних.

Формалізація задачі аналізу потоків даних в онлайн режимі передбачає визначення моделі потоку та алгоритму його обробки [10]. Модель потоку визначає структуру елементів даних, їх взаємозв’язки та характеристики. Алгоритм обробки описує послідовність дій, що виконуються для аналізу потоку та генерації знань. Модель та алгоритм мають враховувати обмеження на використання обчислювальних ресурсів та забезпечувати можливість адаптації до змін характеристик потоку.

Еволюція даних у потоці може відбуватися за різними сценаріями: поступова зміна характеристик, різка зміна характеристик, циклічні зміни, тощо. Розпізнавання типу еволюції та адаптація до нього є завданням

адаптивних алгоритмів, що забезпечують ефективність системи аналізу даних при змінах характеристик потоку. Адаптивність досягається за рахунок моніторингу показників якості моделі та її перенавчання або модифікації при виявленні зниження ефективності.

В контексті кластеризації потоків даних в онлайн режимі застосовується інкрементальний підхід. Інкрементальна кластеризація передбачає поступове оновлення моделі кластерів з надходженням нових елементів даних без необхідності повного перерахунку всіх параметрів моделі. Такий підхід забезпечує обчислювальну ефективність та можливість роботи з потенційно нескінченними потоками даних, які не можуть бути повністю збережені в пам’яті.

Віконний підхід до обробки потоків даних базується на виділенні підмножини елементів потоку, що відповідають певному часовому інтервалу, та їх аналізі. Вікна можуть бути фіксованого розміру (часове вікно) або змінного розміру (адаптивне вікно). Віконний підхід дозволяє обмежити обсяг даних, що обробляються одночасно, та зосередитись на аналізі актуальної інформації, нехтуючи застарілими даними, що можуть не відображати поточний стан досліджуваного процесу.

Концепція забування (forgetting) в аналізі потоків даних передбачає зменшення ваги або повне видалення застарілих елементів даних з моделі [3]. Механізми забування реалізуються через застосування вагових коефіцієнтів, що зменшуються з часом, або через використання віконного підходу з видаленням даних, що виходять за межі вікна. Забування забезпечує адаптивність моделі до змін характеристик потоку та зосередження на аналізі актуальної інформації.

Оцінка якості моделей, побудованих для аналізу потоків даних в онлайн режимі, потребує специфічних підходів. Традиційні методи оцінки, що базуються на порівнянні результатів моделювання з еталонними значеннями, не завжди застосовні через відсутність повної інформації про потік даних та його характеристики. Для оцінки якості моделей в онлайн

режимі використовуються методи прогностичної оцінки, що базуються на порівнянні прогнозованих та фактичних значень, а також методи внутрішньої оцінки, що аналізують структурні характеристики моделі.

Проблема масштабованості в обробці потоків даних пов’язана з необхідністю забезпечення ефективної роботи системи при зростанні обсягу та швидкості надходження даних. Масштабованість досягається через паралелізацію обчислень, розподілену обробку даних, оптимізацію алгоритмів та застосування апроксимаційних методів. Розподілена обробка потоків даних дозволяє розподілити обчислювальне навантаження між кількома вузлами, що забезпечує можливість обробки потоків високої інтенсивності.

Області застосування онлайн аналізу потоків даних охоплюють широкий спектр доменів: моніторинг мережевого трафіку, аналіз поведінки користувачів в мережі Інтернет, обробка даних із сенсорів, фінансові транзакції, тощо [1, 11, 12]. В кожній області застосування потоки даних мають специфічні характеристики, що впливають на вибір алгоритмів та методів їх аналізу. Розуміння особливостей потоків даних в конкретній предметній області дозволяє розробити ефективні алгоритми їх аналізу, що враховують специфіку даних та забезпечують високу якість результатів.

1.2 Методи нечіткої кластеризації: загальна характеристика та класифікація

Нечітка кластеризація є розділом інтелектуального аналізу даних, що займається розподілом елементів множини на підмножини (кластери) з використанням апарату теорії нечітких множин [13]. На відміну від чіткої кластеризації, де кожен елемент однозначно належить до одного кластера, нечітка кластеризація дозволяє елементу належати до кількох кластерів одночасно з різними ступенями належності. Ступінь належності елемента до кластера визначається функцією належності, яка приймає значення в

інтервалі [0, 1], де 0 означає повну відсутність належності, а 1 – повну належність.

Формальна постановка задачі нечіткої кластеризації визначається наступним чином: для множини об’єктів $X = \{x_1, x_2, ..., x_n\}$, де кожен об’єкт представлений вектором ознак у багатовимірному просторі, необхідно побудувати розбиття множини X на с нечітких кластерів $\{A_1, A_2, ..., A_k\}$ таким чином, що $\sum \mu_{ij} = 1$ для всіх $j = 1, 2, ..., n$, де μ_{ij} – ступінь належності j-го об’єкта до i-го кластера. Додатковою умовою є $0 < \sum \mu_{ij} < n$ для всіх $i = 1, 2, ..., c$, тобто жоден кластер не може бути порожнім або містити всі об’єкти з повним ступенем належності.

Методи нечіткої кластеризації базуються на оптимізації цільової функції, яка визначає якість розбиття множини об’єктів на кластери. Цільова функція зазвичай містить компоненти, що відображають компактність кластерів та розділення між ними. Оптимізація цільової функції виконується ітеративними методами, такими як градієнтний спуск, що поступово покращують якість розбиття через модифікацію параметрів кластерів та ступенів належності об’єктів до них.

Одним з фундаментальних методів нечіткої кластеризації є Fuzzy *C*-Means (FCM), який базується на мінімізації цільової функції, що визначає суму зважених відстаней між об’єктами та центрами кластерів [14]. FCM використовує експоненційну вагову функцію з параметром нечіткості $m > 1$, який визначає ступінь нечіткості розбиття. При $m \to 1$ розбиття наближається до чіткого, а при $m \to \infty$ розбиття стає максимально нечітким, з рівними ступенями належності всіх об’єктів до всіх кластерів.

За механізмом формування кластерів методи нечіткої кластеризації поділяються на центроїдні, ієрархічні, сіткові та щільнісні. Центроїдні методи визначають кластери через їх центри та оптимізують розташування цих центрів для мінімізації відстаней між об’єктами та центрами відповідних кластерів. Ієрархічні методи будують дерево вкладених кластерів, де кожен вузол дерева відповідає нечіткому кластеру. Сіткові методи розподіляють

простір ознак на комірки та аналізують належність об’єктів до цих комірок. Щільнісні методи формують кластери як області простору з високою щільністю об’єктів, розділені областями з низькою щільністю.

Possibilistic Fuzzy *C*-Means (PFCM) є модифікацією класичного FCM, яка знімає обмеження на суму ступенів належності. В PFCM ступені належності інтерпретуються як можливості (possibilities) та не обмежуються умовою $\sum \mu_{ij} = 1$, що дозволяє моделювати ситуації, коли об’єкт може бути віднесений до кількох кластерів з високими ступенями належності або не належати до жодного кластера. PFCM є стійким до наявності викидів у даних, оскільки викиди можуть мати низькі ступені належності до всіх кластерів.

Gustafson-Kessel algorithm є узагальненням FCM, що використовує адаптивну метрику відстані на основі нечіткої коваріаційної матриці для кожного кластера. Використання адаптивної метрики дозволяє виявляти кластери довільної форми, а не лише гіперсферичної, як у випадку FCM. Кожен кластер характеризується центром та матрицею, що визначає його форму та орієнтацію у просторі ознак. Алгоритм ітеративно оптимізує розташування центрів кластерів та їх коваріаційні матриці.

Kernel-based Fuzzy *C*-Means використовує ядерні функції для неявного перетворення простору ознак, що дозволяє виявляти нелінійні структури у даних [6]. Ядерна функція визначає міру схожості між об’єктами у неявному просторі ознак вищої розмірності, де лінійні методи кластеризації можуть виявляти нелінійні структури вихідного простору. Вибір ядерної функції (наприклад, радіальної базисної функції, поліноміальної функції) впливає на форму кластерів та результати кластеризації.

Нечіткі нейронні мережі Кохонена поєднують принципи нейронних мереж, що самоорганізуються, та нечіткої логіки для виконання задачі кластеризації [15–18]. Нейрони мережі відповідають прототипам кластерів, а ваги зв’язків – координатам прототипів у просторі ознак. Процес навчання мережі полягає в адаптації ваг зв’язків таким чином, щоб прототипи

розташовувалися в областях з високою щільністю об’єктів. Нечіткість вноситься через визначення функцій належності об’єктів до кластерів на основі відстаней до прототипів.

Інкрементальні методи нечіткої кластеризації призначені для обробки даних, що надходять послідовно, без необхідності повного перерахунку моделі при надходженні нових об’єктів [19]. Інкрементальний FCM модифікує центри кластерів та ступені належності при надходженні нових об’єктів з урахуванням їх впливу на структуру даних. Ефективність інкрементальних методів визначається їх здатністю адаптуватися до змін у структурі даних без втрати якості кластеризації.

Нечіткі методи кластеризації з регуляризацією використовують додаткові терми у цільовій функції для запобігання перенавчанню та отримання більш стабільних результатів. Регуляризація може бути спрямована на забезпечення компактності кластерів, мінімізацію перетинів між кластерами, збереження певної структури даних тощо. Вибір методу регуляризації залежить від конкретної задачі та особливостей аналізованих даних.

Методи нечіткої спектральної кластеризації базуються на аналізі спектра матриці схожості між об’єктами [20]. Спектральна декомпозиція матриці схожості дозволяє виявити приховані структури у даних та виконати кластеризацію у просторі власних векторів матриці. Нечіткість вноситься через використання нечітких методів кластеризації у просторі власних векторів або через визначення нечітких мір схожості між об’єктами.

Оцінка якості нечіткої кластеризації здійснюється з використанням індексів валідності, що враховують специфіку нечіткого розбиття. Індекси валідності можуть враховувати компактність кластерів, їх розділення, однорідність розподілу ступенів належності тощо. Приклади індексів валідності включають Partition Coefficient (PC), Classification Entropy (CE), Xie-Beni index (XB), Fukuyama-Sugeno index (FS). Вибір індексу валідності залежить від особливостей задачі та вимог до якості кластеризації.

Вибір оптимальної кількості кластерів є складною задачею в нечіткій кластеризації, оскільки нечіткі розбиття можуть містити перетинні кластери, що ускладнює визначення їх оптимальної кількості. Методи визначення оптимальної кількості кластерів включають аналіз індексів валідності для різних значень кількості кластерів, використання методів стабільності розбиття, застосування статистичних критеріїв. Автоматичне визначення оптимальної кількості кластерів залишається відкритою проблемою в області нечіткої кластеризації.

Масштабування методів нечіткої кластеризації для роботи з великими обсягами даних потребує модифікації алгоритмів для забезпечення їх обчислювальної ефективності [21]. Підходи до масштабування включають використання випадкових вибірок даних, паралельну обробку, розподілені обчислення, апроксимаційні методи. Ефективне масштабування методів нечіткої кластеризації є необхідною умовою їх практичного застосування для аналізу великих обсягів даних, таких як потоки даних в онлайн режимі.

Методи нечіткої кластеризації мають широкий спектр застосувань: сегментація зображень, класифікація текстів, аналіз медичних даних, моделювання соціальних мереж, тощо [22]. У кожній області застосування методи нечіткої кластеризації адаптуються до специфічних особливостей даних та вимог задачі. Розуміння принципів та обмежень різних методів нечіткої кластеризації дозволяє обрати найбільш відповідний метод для конкретної задачі та досягти оптимальних результатів аналізу даних.

1.3 Принципи правдоподібної кластеризації та її відмінності від традиційних підходів

Правдоподібна кластеризація є методологічним підходом, що ґрунтується на статистичній теорії оцінювання та використовує функцію правдоподібності для формування кластерів. Функція правдоподібності (likelihood function) визначається як ймовірність спостереження наявних

даних за умови заданих параметрів моделі. У контексті кластеризації ця функція виражає ймовірність того, що спостережувані дані належать до конкретного кластера з певними параметрами. Правдоподібна кластеризація спрямована на знаходження таких параметрів моделі, які максимізують функцію правдоподібності для всього набору даних.

Математична формалізація правдоподібної кластеризації спирається на суміш імовірнісних розподілів, де кожен компонент суміші відповідає окремому кластеру [21]. Для набору даних $X = \{x_1, x_2, ..., x_n\}$ та моделі з K кластерами функція правдоподібності визначається як

$$L(\theta \vee X) = \prod_{i=1}^{n} f(x_i \vee \theta), \quad (1.1)$$

де $f(x_i|\theta) = \sum_{j=1}^{k} \pi_j \cdot f_j(x_i|\theta_j)$.

π_j – апріорна ймовірність належності до j-го кластера (вага компонента)

$f_j(x_i|\theta_j)$ – функція щільності ймовірності j-го компонента з параметрами θ_j.

Задача кластеризації полягає у знаходженні параметрів θ, що максимізують функцію правдоподібності.

Основна відмінність правдоподібної кластеризації від традиційних підходів полягає у її статистичній інтерпретації [23]. Традиційні методи, такі як K-means, базуються на геометричних концепціях відстані та подібності об’єктів, тоді як правдоподібна кластеризація інтерпретує формування кластерів як процес статистичного моделювання даних. Це дозволяє враховувати невизначеність та шум у даних, а також надає формальні критерії для вибору оптимальної моделі та кількості кластерів.

Expectation-Maximization (EM) алгоритм є основним методом оптимізації в правдоподібній кластеризації . EM алгоритм складається з двох кроків: кроку очікування (E-step), на якому обчислюються ступені належності об’єктів до кластерів на основі поточних параметрів моделі, та кроку максимізації (M-step), на якому оновлюються параметри моделі для максимізації функції правдоподібності. Ці кроки повторюються ітеративно

до досягнення збіжності, що гарантує знаходження локального максимуму функції правдоподібності.

Правдоподібна нечітка кластеризація поєднує принципи правдоподібності та нечіткої логіки, дозволяючи об’єктам належати до кількох кластерів одночасно з різними ступенями належності. Нечіткість вноситься через розгляд ступенів належності як апостеріорних ймовірностей належності об’єктів до кластерів, обчислених за формулою Баєса. Ці ймовірності використовуються як ваги при обчисленні параметрів кластерів на кроці максимізації ЕМ алгоритму. Таким чином, правдоподібна нечітка кластеризація надає імовірнісну інтерпретацію нечіткому розбиттю даних.

Гауссові суміші (Gaussian Mixture Models, GMM) є найбільш поширеною моделлю в правдоподібній кластеризації [24]. У GMM кожен кластер моделюється багатовимірним нормальним розподілом з параметрами μ_k (вектор середніх) та Σ_k (коваріаційна матриця). Функція щільності ймовірності для об’єкта x_i та кластера k визначається як

$$f_{\mathrm{k}}(x_i \vee \mu_k, \Sigma_k) = (2\pi)^{-d/\ 2} \vee \Sigma_k \vee^{-\ 1/\ 2} exp\left(-½(x_i - \mu_k)^{\mathrm{T}} \Sigma_k^{-1(x_{\mathrm{i}} - \mu_{\mathrm{k}})}\right), (1.2)$$

де d – розмірність простору ознак.

GMM є гнучкою моделлю, що може апроксимувати складні розподіли даних та виявляти кластери довільної форми.

Вибір оптимальної кількості компонентів у суміші є однією з проблем правдоподібної кластеризації. Збільшення кількості компонентів завжди підвищує значення функції правдоподібності, але може призвести до перенавчання моделі. Для вирішення цієї проблеми використовуються інформаційні критерії, такі як AIC (Akaike Information Criterion) та BIC (Bayesian Information Criterion), що балансують між точністю моделі та її складністю. Ці критерії мають вигляд AIC = $-2ln(L) + 2p$ та BIC = $-2ln(L) + p \cdot ln(n)$, де L – максимальне значення функції правдоподібності, p – кількість параметрів моделі, n – розмір вибірки.

Робастна правдоподібна кластеризація спрямована на забезпечення стійкості до викидів та шуму в даних. Робастність досягається через модифікацію функції правдоподібності або через використання сумішей розподілів з «важкими хвостами», таких як *t*-розподіл. Інший підхід полягає у введенні додаткового фонового компонента у суміш, який моделює шум та викиди. Робастні методи правдоподібної кластеризації є особливо актуальними для аналізу реальних даних, які часто містять шум та аномалії.

Онлайн версія правдоподібної кластеризації адаптує принципи максимізації правдоподібності для роботи з потоками даних в реальному часі [3]. Основною проблемою є необхідність інкрементального оновлення параметрів моделі без повторної обробки всіх історичних даних. Онлайн EM алгоритм використовує стохастичну апроксимацію для оновлення параметрів моделі з надходженням нових даних. Для кожного нового спостереження x_t обчислюються апостеріорні ймовірності його належності до кластерів, і параметри моделі оновлюються з використанням цих ймовірностей та коефіцієнта навчання, що залежить від часу.

Механізми забування в правдоподібній кластеризації потоків даних реалізуються через введення вагових коефіцієнтів для спостережень. Ваги зменшуються з часом, що призводить до зменшення впливу застарілих даних на параметри моделі. Формально це можна представити як модифікацію функції правдоподібності:

$$L(\theta|X) = \prod \prod_{i=1}^{n} [f(x_i|\theta)]^{w_i}, \qquad (1.3)$$

де w_i – вага i-го спостереження, що залежить від його віку.

Експоненційне забування використовує ваги виду $w_i = \lambda^{(t-t_i)}$, де $\lambda \in (0, 1)$ – параметр забування, t – поточний час, t_i – час надходження i-го спостереження. Механізми забування забезпечують адаптивність моделі до змін у структурі даних з плином часу.

Правдоподібна кластеризація надає природний механізм для виявлення та обробки аномалій у потоках даних [25]. Аномалії визначаються як спостереження з низькою правдоподібністю відносно всіх компонентів суміші. Формально, спостереження x_i вважається аномалією, якщо $f(x_i|\theta) < \tau$, де τ – деяке порогове значення. Онлайн версія правдоподібної кластеризації дозволяє виявляти аномалії в режимі реального часу та адаптувати модель для врахування нових типів нормальної поведінки, що може виникати у процесі еволюції даних.

Порівняння правдоподібної кластеризації з традиційними методами, такими як FCM та *K*-means, демонструє її переваги у статистичній інтерпретації результатів та можливості моделювання складних розподілів даних [14]. FCM та *K*-means базуються на геометричних концепціях відстані та мінімізації внутрішньокластерних дисперсій, що обмежує їх здатність виявляти кластери складної форми. Правдоподібна кластеризація, використовуючи суміші розподілів, може моделювати кластери довільної форми та враховувати кореляції між ознаками через коваріаційні матриці компонентів суміші.

Інтеграція правдоподібної кластеризації з еволюційними алгоритмами відкриває нові можливості для оптимізації параметрів моделі в умовах потоків даних. Еволюційні алгоритми можуть бути використані для пошуку глобального максимуму функції правдоподібності, уникаючи локальних оптимумів, у які може потрапити EM алгоритм. Наприклад, генетичні алгоритми або алгоритми котячих зграй можуть оптимізувати параметри суміші розподілів, використовуючи функцію правдоподібності як цільову функцію. Це особливо актуально для складних моделей з великою кількістю параметрів, де простір пошуку має складну структуру.

Оцінка якості правдоподібної кластеризації потоків даних потребує спеціалізованих підходів. Традиційні індекси валідності, розроблені для статичних даних, не враховують темпоральний аспект потоків даних та зміни їх структури з часом. Для оцінки якості онлайн версії правдоподібної

кластеризації використовуються прогностичні індекси, що аналізують здатність моделі передбачати розподіл нових даних. Крім того, особлива увага приділяється оцінці адаптивності моделі до змін у структурі даних, що може бути виражена через швидкість відновлення якості кластеризації після зміни характеристик потоку.

Прикладні аспекти правдоподібної нечіткої кластеризації потоків даних охоплюють різноманітні домени: аналіз соціальних мереж, моніторинг мережевого трафіку, обробка фінансових транзакцій, тощо [3]. У кожному домені правдоподібна кластеризація адаптується до специфічних характеристик даних та вимог задачі. Наприклад, у аналізі соціальних мереж правдоподібна кластеризація може бути використана для виявлення спільнот користувачів з подібними інтересами та моніторингу еволюції цих спільнот з часом. У моніторингу мережевого трафіку правдоподібна кластеризація дозволяє виявляти типові паттерни трафіку та аномалії, що можуть вказувати на мережеві атаки.

Інтеграція онтологій з правдоподібною нечіткою кластеризацією потоків даних відкриває нові можливості для інтерпретації результатів кластеризації. Онтології визначають семантичні взаємозв’язки між концепціями предметної області, що дозволяє збагатити результати кластеризації додатковими знаннями та забезпечити їх інтерпретабельність. Наприклад, у аналізі текстових потоків даних онтології можуть бути використані для визначення семантичної близькості між термінами, що дозволяє виявляти кластери текстів, пов’язаних з певними поняттями, навіть якщо ці тексти не містять однакових слів [26]. Інтеграція онтологій з правдоподібною кластеризацією реалізується через модифікацію функції правдоподібності для врахування семантичних взаємозв’язків між об’єктами.

Паралельні та розподілені реалізації правдоподібної нечіткої кластеризації потоків даних спрямовані на забезпечення масштабованості та ефективності обробки великих обсягів даних [20, 27]. Паралельні реалізації EM алгоритму розподіляють обчислення між кількома процесорами або

ядрами, що дозволяє прискорити обробку даних. Розподілені реалізації використовують парадигму MapReduce для обробки даних на кластері обчислювальних вузлів. Кожен вузол обробляє частину даних, обчислюючи локальні параметри моделі, які потім агрегуються для отримання глобальних параметрів. Такі реалізації забезпечують можливість обробки потоків даних високої інтенсивності та складності.

Перспективними напрямками розвитку правдоподібної нечіткої кластеризації потоків даних є інтеграція з глибоким навчанням, розробка методів для обробки неструктурованих та мультимодальних даних, вдосконалення механізмів адаптації до змін у структурі даних. Глибокі нейронні мережі можуть бути використані для автоматичного вилучення інформативних ознак з сирих даних, що підвищує якість подальшої кластеризації [15, 28–31]. Методи для обробки неструктурованих та мультимодальних даних дозволяють застосовувати правдоподібну кластеризацію до широкого спектру задач, таких як аналіз текстів, зображень, відео, тощо. Вдосконалені механізми адаптації забезпечують стабільну роботу системи кластеризації в умовах динамічно змінюваних характеристик потоку даних.

1.4 Постановка задачі

Метою роботи є розробка та моделювання методу правдоподібної нечіткої кластеризації потоків даних, що забезпечує ефективний аналіз неперервних послідовностей спостережень в онлайн режимі з урахуванням невизначеності та статистичної природи даних.

Для досягнення цієї мети поставлено такі завдання:

– провести аналіз існуючих методів нечіткої кластеризації та їх адаптацій для роботи з потоками даних в онлайн режимі;

– розробити математичну модель методу правдоподібної нечіткої кластеризації потоків даних;

– реалізувати програмну модель розробленого методу та провести експериментальні дослідження його ефективності;

– провести порівняльний аналіз запропонованого методу з існуючими підходами до кластеризації потоків даних.

Об’єктом роботи є процеси кластеризації потоків даних в онлайн режимі з використанням методів нечіткої логіки та статистичної теорії оцінювання.

Предметом роботи є методи, моделі та алгоритми правдоподібної нечіткої кластеризації потоків даних в онлайн режимі.

Робота пов’язана з іншими роботами в галузі інтелектуального аналізу даних, зокрема з розробками в області адаптивних методів нечіткої кластеризації та статистичних підходів до аналізу потоків даних.

2 МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ МЕТОДУ ПРАВДОПОДІБНОЇ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ

2.1 Формалізація задачі нечіткої кластеризації потоків даних

Формалізація задачі нечіткої кластеризації потоків даних базується на розширенні класичної постановки задачі кластеризації з урахуванням двох специфічних аспектів: нечіткої природи приналежності об'єктів до кластерів та динамічного характеру потоків даних. У контексті потоків даних розглядається нескінченна послідовність спостережень $X = \{x_1, x_2, ..., x_n, ...\}$, де кожне спостереження x_t є вектором ознак у d-вимірному просторі R^d, що надходить у момент часу t. Задача полягає у визначенні c нечітких кластерів та обчисленні для кожного спостереження x_t вектора значень функцій приналежності $\mu(x_t) = [\mu_1(x_t), \mu_2(x_t), ..., \mu c(x_t)]^T$, де $\mu_i(x_t) \in [0, 1]$ визначає ступінь приналежності спостереження x_t до i-го кластера.

Математична формалізація нечітких кластерів у потоці даних ґрунтується на використанні нечітких множин. Кожен кластер A_i, $i = 1, 2, ..., c$, представляється як нечітка множина в просторі ознак R^d, що характеризується функцією приналежності μ_i: $R^d \rightarrow [0, 1]$. Функція приналежності визначає ступінь, з яким кожне спостереження у потоці даних належить до відповідного кластера. Сукупність таких функцій для всіх c кластерів утворює нечітке розбиття простору ознак [24].

Для формалізації обмежень на функції приналежності у задачі нечіткої кластеризації потоків даних використовуються наступні умови: $\sum_{i=1}^{c} \mu_i(x_t) = 1$ для всіх t, що забезпечує нормалізацію ступенів приналежності; $0 < \sum_{t=1}^{T} \mu_i(x_t) < T$ для всіх $i = 1, 2, ..., c$, де T – кількість спостережень у розглянутому часовому вікні, що гарантує відсутність порожніх кластерів та забезпечує розподіл спостережень між різними кластерами.

Часова складова у формалізації задачі нечіткої кластеризації потоків даних враховується через введення часової залежності для параметрів

кластерів. Кожен кластер A_i характеризується центром $v_i(t)$ та, можливо, додатковими параметрами, такими як матриця коваріації $\Sigma_i(t)$ або радіус $r_i(t)$, які змінюються з часом. Еволюція цих параметрів описується диференціальними рівняннями або різницевими схемами, що враховують надходження нових спостережень та механізми забування застарілих даних.

Механізми забування у формалізації задачі базуються на введенні вагових коефіцієнтів для спостережень залежно від їх віку. Експоненційна схема забування визначає вагу спостереження x_t у момент часу $\tau > t$ як $w(x_t, \tau) = \lambda^{\tau-t}$, де $\lambda \in (0, 1)$ – параметр забування. Лінійна схема визначає вагу як

$$w(x_\mathrm{t}, \tau) = max\left(0{,}1 - \frac{\tau - t}{\Delta}\right), \qquad (2.1)$$

де Δ – параметр, що визначає максимальний вік спостережень, які враховуються. Ці механізми забезпечують адаптивність моделі до змін у структурі даних з часом.

В умовах онлайн режиму роботи, формалізація задачі нечіткої кластеризації потоків даних повинна враховувати обчислювальні обмеження: алгоритм має обробляти кожне спостереження x_t за фіксований час, незалежно від кількості раніше оброблених спостережень; обсяг пам’яті, що використовується алгоритмом, має бути обмеженим та не залежати від загальної кількості спостережень у потоці. Ці обмеження зумовлюють необхідність розробки інкрементальних алгоритмів, здатних оновлювати модель без повного перерахунку всіх параметрів.

Таблиця 2.1 – Порівняння формалізацій задач нечіткої кластеризації для статичних даних та потоків даних

Характеристика	**Статичні дані**	**Потоки даних**
Розмір набору даних	Фіксований (n)	Потенційно нескінченний
Доступ до даних	Багаторазовий доступ до всього набору	Однократний доступ до кожного елемента
Часова складова	Не враховується	Явно моделюється
Обчислювальна складність	O(n) допустима	Вимагається O(1) для кожного нового елемента
Вимоги до пам’яті	O(n) допустима	Вимагається O(1) незалежно від тривалості потоку
Параметри кластерів	Статичні протягом всього процесу	Динамічно змінюються з часом
Механізми забування	Не застосовуються	Необхідні для адаптації до змін

Цільова функція у формалізації задачі нечіткої кластеризації потоків даних узагальнює класичну цільову функцію Fuzzy *C*-Means для врахування часової складової та механізмів забування. У загальному вигляді вона може бути представлена як

$$J(\mu, v, t) = \sum_{k=1}^{K}\sum_{i=1}^{T} c^{w}(x_k, t) \cdot [\mu_{\mathrm{i}}(x_{\mathrm{k}})]^{\mathrm{m}} \cdot d^2(x_k, v_i(t)), \qquad (2.2)$$

де $m > 1$ – параметр нечіткості;

$d(\cdot, \cdot)$ – функція відстані;

$w(x_k, t)$ –вагова функція, що реалізує механізм забування.

Мінімізація цієї функції відносно функцій приналежності μ_i та параметрів кластерів v_i з урахуванням обмежень є математичною постановкою задачі нечіткої кластеризації потоків даних [14].

Формалізація адаптивного механізму визначення кількості кластерів є суттєвим аспектом нечіткої кластеризації потоків даних. У динамічному середовищі кількість кластерів може змінюватися з часом через зміни

структури даних. Формальне представлення цього механізму включає критерії для розщеплення існуючого кластера (наприклад, якщо його внутрішня дисперсія перевищує пороговий рівень) та для об'єднання кластерів (наприклад, якщо відстань між їх центрами стає меншою за пороговий рівень).

Якість нечіткої кластеризації потоків даних оцінюється з використанням адаптованих індексів валідності. Для часового вікна [t-Δ, t] індекс нечіткого розбиття може бути визначений як

$$PC(t) = \left(\frac{1}{|W(t)|}\right) \cdot \sum_{x_k} \in W(t) \sum_{i=1}^{c} [\mu_{\text{i}}(x_{\text{k}})]^2, \qquad (2.3)$$

де $W(t)$ – множина спостережень у часовому вікні;

$|W(t)|$ – їх кількість.

Індекс ентропії класифікації визначається як $CE(t) = -\left(\frac{1}{|W(t)|}\right) \cdot \sum_{x_k} \in W(t) \sum_{i=1}^{c} \mu_{\text{i}}(x_{\text{k}}) \cdot log\left(\mu_{\text{i}}(x_{\text{k}})\right)$. Ці індекси дозволяють оцінювати якість кластеризації в режимі реального часу та використовуються для адаптації параметрів алгоритму.

Проблема дрейфу концепцій (concept drift) формалізується через зміни розподілу даних з часом. У контексті кластеризації це може проявлятися як зміни кількості, форми, розміру та розташування кластерів. Формально, дрейф концепцій описується як зміна спільного розподілу вхідних ознак та прихованої структури кластерів $P(X, Y)$ з часом. Виявлення та адаптація до дрейфу концепцій є критичними аспектами нечіткої кластеризації потоків даних в онлайн режимі [3, 32].

У формалізації проблеми викидів та шуму в потоках даних застосовуються спеціальні механізми в моделі нечіткої кластеризації. Один підхід передбачає створення додаткового фонового кластера A_0 з власною функцією приналежності μ_0. Ця функція кількісно визначає ступінь, з яким певне спостереження класифікується як шум або викид. Інший метод

зосереджується на зміні цільової функції, щоб зменшити вплив віддалених спостережень на обчислення параметрів кластерів.

Багатомасштабна формалізація розширює можливості нечіткої кластеризації потоків даних через аналіз на різних часових рівнях. Для кожного масштабу s (від 1 до S) встановлюється індивідуальний механізм забування з відповідним параметром λ_s. Це дозволяє отримати специфічний для конкретного масштабу набір кластерів $\{A_1^s, A_2^s, ..., Ac_s^s\}$, кожен з унікальними функціями приналежності.

Особливість такого підходу полягає в можливості комбінувати результати кластеризації з різних часових масштабів. Завдяки цьому система може одночасно виявляти патерни різної тривалості – від короткострокових до довгострокових – забезпечуючи повніше розуміння структури потоку даних.

Ця методологія дозволяє ефективно працювати з неоднорідностями в даних, зберігаючи точність кластеризації навіть за наявності шуму. Багатомасштабний аналіз також підвищує адаптивність системи до змін у часовій динаміці потоків даних, що робить його цінним інструментом для різноманітних практичних застосувань [33, 34].

Формалізація інкрементальної нечіткої кластеризації потоків даних базується на рекурентних формулах для оновлення параметрів кластерів. Для онлайн версії Fuzzy *C*-Means це можна представити як

$$v_{\mathrm{i}}(t) = v_{\mathrm{i}}(t-1) + \eta(t) \cdot \mu_i^m(x_{\mathrm{t}}) \cdot \frac{\left(x_{\mathrm{t}} - v_{\mathrm{i}}(t-1)\right)}{\left(S_{\mathrm{i}}(t-1) + \mu_i^m(x_{\mathrm{t}})\right)}, \qquad (2.4)$$

де $S_i(t) = \lambda \cdot S_i(t\text{-}1) + \mu_i^m(x_i)$ – накопичена вага і-го кластера;

$\eta(t)$ – коефіцієнт навчання, що може залежати від часу.

Аналогічні рекурентні формули можуть бути розроблені для інших параметрів кластерів, таких як матриці коваріації [21, 35].

Формалізація розподіленої нечіткої кластеризації потоків даних є актуальною для систем з множинними джерелами даних. У цьому випадку визначаються локальні моделі кластеризації для кожного джерела, які періодично агрегуються для отримання глобальної моделі. Формалізація включає визначення стратегій агрегації локальних параметрів кластерів (наприклад, зважене усереднення центрів кластерів) та механізмів синхронізації локальних моделей з глобальною.

2.2 Функції правдоподібності у задачах кластерного аналізу

Функції правдоподібності у задачах кластерного аналізу є основним інструментом статистичного підходу до формування та оцінки моделей кластеризації. Функція правдоподібності визначається як ймовірність спостереження наявних даних за умови заданих параметрів моделі. Для набору даних $X = \{x_1, x_2, ..., x_n\}$ та параметрів моделі θ функція правдоподібності має вигляд $L(\theta|X) = P(X|\theta) = \prod^{n}_{i=1} P(x_i|\theta)$, де $P(x_i|\theta)$ – ймовірність спостереження x_i за умови параметрів моделі θ. У контексті кластеризації параметри моделі включають параметри розподілів, що описують кластери, та апріорні ймовірності належності до кластерів [36].

У моделі суміші розподілів, що становить основу правдоподібної кластеризації, кожен кластер представляється окремим компонентом суміші з властивим йому розподілом. Функція щільності ймовірності в такій моделі визначається як $f(x|\theta) = \sum^{k}_{j=1} \pi_j \cdot f_j(x|\theta_j)$, де π_j позначає апріорну ймовірність належності до j-го кластера (вага компонента), а $f_j(x|\theta_j)$ – функцію щільності ймовірності j-го компонента з параметрами θ_j. Для всього набору даних функція правдоподібності обчислюється як $L(\theta|X) = \prod^{n}_{i=1} \sum^{k}_{j=1} \pi_j \cdot f_j(x_i|\theta_j)$. Процес максимізації цієї функції стосовно параметрів θ становить сутність задачі правдоподібної кластеризації.

На практиці замість звичайної функції правдоподібності застосовується її логарифмічна форма, що дозволяє спростити обчислення та підвищити

чисельну стабільність. Ця форма подається як $\ell(\theta|X) = log\ L(\theta|X) = \sum^{n}_{i=1} log\ (\sum^{k}_{j=1} \pi_j \cdot f_j(x_i|\theta_j))$. Максимізація логарифмічної функції правдоподібності дає такий самий результат, як і максимізація звичайної функції, проте є зручнішою з обчислювальної перспективи. Слід зазначити, що наявність суми всередині логарифма створює певні труднощі для аналітичного розв'язання задачі максимізації.

В ЕМ-алгоритмі, що застосовується для правдоподібної кластеризації, використовується очікувана повна логарифмічна функція правдоподібності. Цей підхід ґрунтується на введенні прихованих змінних z_{ij}, які визначають приналежність спостереження x_i до кластера j: якщо $z_{ij} = 1$, то спостереження належить до кластера, якщо $z_{ij} = 0$ – не належить. Повна функція правдоподібності формулюється як

$$L_{c(\theta|X,Z)} = \prod_{i=1}^{n}\prod_{j=1}^{n}\left[\pi_j \cdot f_j(x_i|\theta_j)\right]^{z_{ij}}, \tag{2.5}$$

а її логарифм – як

$$logL_{c(\theta|X,Z)} = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \cdot \left[log\pi_j + logf_j(x_i|\theta_j)\right], \tag{2.6}$$

Оскільки змінні Z безпосередньо не спостерігаються, ЕМ-алгоритм оперує очікуваним значенням l_c відносно умовного розподілу Z за заданих X та поточних параметрах θ.

Гауссівська функція правдоподібності є найбільш поширеною у правдоподібній кластеризації. Вона базується на моделі, де кожен кластер описується багатовимірним нормальним розподілом. Функція щільності ймовірності для такого розподілу має вигляд

$$f_j(x|\mu_j, \Sigma_j) = (2\pi)^{\frac{-d}{2}}|\Sigma_j|^{\left(\frac{-1}{2}\right)exp}\left(-\left(\frac{1}{2}\right)(x-\mu_j)^T\Sigma_j^{(-1)(x-\mu_j)}\right), \tag{2.7}$$

де μ_j – вектор середніх значень;

Σ_j – коваріаційна матриця;

d – розмірність даних.

Відповідна функція правдоподібності для моделі суміші гауссіанів (GMM) є добутком сум таких функцій щільності ймовірності, зважених на відповідні π_j.

Таблиця 2.2 – Типи функцій правдоподібності, що використовуються в задачах кластерного аналізу

Тип функції правдоподіб-ності	**Математичний вираз**	**Характеристи-ки**	**Застосування**
Гауссівська	$f_{\mathrm{j}}(x\|\mu_{\mathrm{j}},\Sigma_{\mathrm{j}}) = (2\pi)^{\frac{-d}{2}}\|\Sigma_{\mathrm{j}}\|^{\left(\frac{-1}{2}\right)exp}\left(-\left(\frac{1}{2}\right)(x-\mu_{\mathrm{j}})^T\Sigma_{\mathrm{j}}^{(-1)(x-\mu_{\mathrm{j}})}\right)$	Симетричний розподіл з «легкими хвостами»	Загальна кластеризація з еліптичними кластерами
t-розподіл	$f_{\mathrm{j}}(x\|\mu_{\mathrm{j}},\Sigma_{\mathrm{j}},\nu_{\mathrm{j}}) \propto \left[1+\left(\frac{1}{\nu_{\mathrm{j}}}\right)(x-\mu_{\mathrm{j}})^T\sum_j^{-1}(x-\mu_j)\right]^{\frac{-\nu_{\mathrm{j}}+d}{2}}$	Симетричний з «важчими хвостами», параметр ν_j контролює «важкість хвостів»	Робастна кластеризація за наявності викидів
Бернуллі	$f_{\mathrm{j}}(x\|p_{\mathrm{j}}) = \prod_k p_{jk}^{x_k}(1-p_{\mathrm{jk}})^{1-x_{\mathrm{k}}}$	Для бінарних даних, $p_{j\mathrm{k}}$ - ймовірність 1 для k-ї ознаки у j-му кластері	Кластеризація бінарних даних
Поліноміаль-на	$f_{\mathrm{j}}(x\|p_{\mathrm{j}}) = \frac{n!}{\prod_k x_{\mathrm{k}}!}\cdot\prod_k p_{jk}^{x_k}$	Для даних з підрахунком, p_{jk} - ймовірність k-ї категорії у j-му кластері	Кластеризація текстових даних (моделі «мішок слів»)
Експоненційн-на	$f_j(x\|\lambda_j) = \lambda_j exp(-\lambda_j x)$ для $x\geq 0$	Для невід'ємних даних, λ_j – параметр масштабу	Аналіз часових інтервалів, тривалості подій

Функції правдоподібності на основі *t*-розподілу використовуються для робастної кластеризації, стійкої до викидів. *t*-розподіл має «важчі хвости» порівняно з нормальним розподілом, що дозволяє зменшити вплив віддалених спостережень на параметри кластерів. Функція щільності ймовірності для *t*-розподілу має додатковий параметр ν, що визначає кількість ступенів свободи: $f_j(x|\mu_j, \Sigma_j, \nu_j) \propto \left[1 + \left(\frac{1}{\nu_j}\right)(x-\mu_j)^T \Sigma_j^{(-1)(x-\mu_j)}\right]^{\frac{-\nu_j+d}{2}}$. При $\nu_j \to \infty$ *t*-розподіл наближається до нормального, а при малих значеннях ν_j хвости розподілу стають важчими, що забезпечує робастність [14].

При роботі з категоріальними даними функції правдоподібності спираються на поліноміальні або мультиноміальні розподіли. Розподіл Бернуллі застосовується переважно для бінарних даних, тоді як поліноміальний розподіл використовується для даних з підрахунком. Функція правдоподібності для моделі суміші поліноміальних розподілів математично виражається як $L(\theta|X) = \prod^{n}_{i=1} \sum^{k}_{j=1} \pi_j \cdot \prod^{d}_{k=1} p_{jk}\wedge x_{ik}$. У цьому виразі p_{jk} означає ймовірність спостереження *k*-ї категорії в *j*-му кластері. Ця модельна конструкція знаходить широке застосування при кластеризації текстових даних, де x_{ik} відображає частоту появи *k*-го слова в *i*-му документі.

Для нечіткої кластеризації даних з експоненційним розподілом, що характеризує тривалість подій або часових інтервалів, застосовуються експоненційні функції правдоподібності. Функція щільності ймовірності для одновимірного експоненційного розподілу представляється як $f_j(x|\lambda_j) = \lambda_j exp(-\lambda_j x)$ за умови $x \geq 0$, де $\lambda_j > 0$ виступає параметром розподілу. При переході до багатовимірного випадку використовуються різноманітні узагальнення експоненційного розподілу, вибір яких залежить від припущень про взаємозв’язки між змінними.

Мішані функції правдоподібності знаходять своє застосування при спільній кластеризації даних різнотипного характеру. Для даних, що поєднують неперервні та категоріальні ознаки, функція правдоподібності

може виражатися як добуток гауссівських компонент (для неперервних ознак) та поліноміальних компонент (для категоріальних):

$$f_j(x|\theta_j) = f_{jcont}(x_{cont}|\mu_j, \Sigma_j) \cdot f_{jcat}(x_{cat}|p_j), \quad (2.8)$$

У цьому виразі x_{cont} і x_{cat} позначають неперервні та категоріальні складові вектора ознак відповідно. Цей підхід дозволяє враховувати гетерогенну природу даних при виконанні кластеризації.

Байєсівський підхід до функцій правдоподібності в кластерному аналізі передбачає використання апріорних розподілів для параметрів моделі θ та обчислення апостеріорного розподілу $p(\theta|X) \propto L(\theta|X) \cdot p(\theta)$. Такий метод надає можливість інтегрувати апріорні знання про структуру даних та забезпечує регуляризацію моделі. Це набуває особливого значення при роботі з обмеженими наборами даних або при наявності шумових компонентів. У випадку гауссівських сумішей зазвичай використовуються такі апріорні розподіли: нормальний розподіл для векторів середніх значень та розподіл Уішарта для коваріаційних матриць.

Застосування різних функцій правдоподібності дозволяє адаптувати методи кластеризації до специфіки аналізованих даних, забезпечуючи більш точний аналіз прихованих структур. Моделювання кластерів з використанням відповідних статистичних розподілів сприяє підвищенню ефективності процесу кластеризації та формуванню обґрунтованих висновків щодо структурних особливостей даних.

Правдоподібність для нечіткої кластеризації поєднує класичну функцію правдоподібності з елементами нечіткої логіки. Нечіткість у цьому контексті може бути інтерпретована через ймовірнісні вагові коефіцієнти для спостережень. Функцію правдоподібності для нечіткої моделі можна представити як

$$L(\theta|X,\mu) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[f_{\mathrm{j}}(x_{\mathrm{i}}|\theta_{\mathrm{j}})\right]^{\mu_{\mathrm{ij}}}, \quad (2.9)$$

де $\mu_{ij} \in [0, 1]$ – ступінь належності спостереження x_i до кластера j, що задовольняє умові $\sum_{j=1}^{k} \mu_{ij} = 1$ для всіх i.

Ця форма функції правдоподібності є узагальненням класичної функції для моделі суміші, коли μ_{ij} інтерпретуються як апостеріорні ймовірності належності до кластерів [24].

Для потоків даних функція правдоподібності модифікується для врахування часової складової та механізмів забування. Часова правдоподібність може бути представлена як

$$L\big(\theta(t) \vee X(t)\big) = \prod_{k=1}^{t} w(k,t) \cdot \sum_{j=1}^{K} \pi_j(t) \cdot f_{\mathrm{j}}\left(x_{\mathrm{k}} \vee \theta_{\mathrm{j}}(t)\right), \quad (2.10)$$

де $w(k, t)$ – вагова функція, що реалізує механізм забування і зменшує вплив застарілих спостережень,

$\theta(t)$ – параметри моделі в момент часу t;

$X(t)$ – множина всіх спостережень до моменту t;

Ця форма функції правдоподібності дозволяє адаптувати модель до змін у структурі даних з часом.

Метод максимальної правдоподібності (Maximum Likelihood Estimation, MLE) використовується для знаходження оптимальних значень параметрів θ, які максимізують функцію правдоподібності $L(\theta|X)$ або її логарифмічний еквівалент $\ell(\theta|X)$. При роботі з моделями сумішей розподілів аналітичне рішення стає недоступним через структурні особливості – наявність суми всередині логарифма. Саме тому застосовуються ітераційні методи, серед яких виділяється EM-алгоритм.

В умовах онлайн кластеризації потоків даних класичні методи потребують адаптації. Інкрементальні версії алгоритмів дозволяють оновлювати параметричні оцінки при надходженні нових спостережень,

уникаючи необхідності повного перерахунку моделі. Такий підхід забезпечує ефективну обробку потокових даних без накопичення всієї історії спостережень.

Вибір оптимальної кількості компонентів у моделі суміші здійснюється за допомогою інформаційних критеріїв. AIC (критерій Акаіке) та BIC (байєсівський інформаційний критерій) збалансовують точність моделі та її складність. Формально ці критерії представляються наступним чином:

$$AIC = -2l(\theta|X) + 2pBIC = -2l(\theta|X) + p \cdot log(n), \qquad (2.11)$$

де θ – позначає оцінку параметрів за методом максимальної правдоподібності;

p – відображає кількість параметрів моделі;

n – відповідає розміру вибірки.

При порівнянні моделей перевага надається тій, що має менше значення AIC або BIC [21].

Виявлення викидів у даних також може спиратись на функцію правдоподібності. Базовий підхід ґрунтується на припущенні, що спостереження-викиди характеризуються низькою ймовірністю відносно всіх компонентів суміші. Формально, спостереження x_i класифікується як викид, якщо:

$$P(x_\text{i} \vee \theta) = \sum_{j=1}^{k} \pi_j \cdot f_\text{j}(x_\text{i} \vee \theta_\text{j}) < \tau, \qquad (2.12)$$

де τ представляє деякий пороговий рівень.

Існує також альтернативний метод обробки викидів через введення додаткового компонента в модель суміші. Цей компонент моделює рівномірний розподіл у просторі ознак і фактично «захоплює» спостереження-викиди. При такому підході функція правдоподібності модифікується до вигляду:

$$P(x_\text{i} \vee \theta) = (1 - \pi^0) \cdot \sum_{j=1}^{k} \pi_j \cdot f_\text{j}(x_\text{i} \vee \theta_\text{j}) + \pi^0 \cdot U(x_\text{i}), \qquad (2.13)$$

де π_0 позначає апріорну ймовірність того, що спостереження є викидом;

$U(x_i)$ – відповідає щільності рівномірного розподілу.

Такий підхід дозволяє не лише ідентифікувати викиди в даних, але й зменшити їх вплив на оцінку параметрів основних компонентів моделі, забезпечуючи більш стійку кластеризацію в умовах зашумлених даних та аномалій.

2.3 Розробка алгоритму правдоподібної нечіткої кластеризації для онлайн обробки

Алгоритм правдоподібної нечіткої кластеризації для онлайн обробки потоків даних поєднує статистичний підхід на основі функції правдоподібності з принципами нечіткої логіки. Така комбінація дозволяє об’єктам одночасно належати до декількох кластерів з різним ступенем належності. Розроблений алгоритм має відповідати вимогам обчислювальної ефективності, вміти адаптуватися до змін структури даних, демонструвати стійкість до шуму й викидів та функціонувати при обмежених ресурсах пам’яті.

В основі алгоритму лежить інкрементальна версія методу очікування-максимізації (Online Expectation-Maximization). Під час ініціалізації встановлюється початкова кількість кластерів c та визначаються їхні параметри. Для кожного кластера j (де j = 1, 2, ..., c) задаються:

- вектор середніх значень $\mu_j(0)$;
- коваріаційна матриця $\Sigma_j(0)$;
- апріорна ймовірність $\pi_j(0)$.

Для гауссівської моделі ці параметри встановлюються на основі перших спостережень потоку або заздалегідь визначених значень, що враховують апріорні знання про структуру даних [36, 37].

Головний цикл алгоритму обробляє кожне нове спостереження x_t, що надходить у момент часу t, та оновлює параметри моделі. Перший крок циклу – обчислення нечітких ступенів належності спостереження до кластерів, які розглядаються як апостеріорні ймовірності:

$$\mu_{\mathrm{j}}(x_{\mathrm{t}}) = \pi_{\mathrm{j}}(t-1) \cdot f_{\mathrm{j}}(x_t \vee \theta_j(t-1)) / \sum_{i=1}^{k} \pi_i(t-1) \cdot f_{\mathrm{i}}\left(x_{\mathrm{t}} \middle| \theta_{\mathrm{i}}(t-1)\right). \quad (2.14)$$

У випадку гауссівської моделі функція $f_j(x_t|\theta_j(t\text{-}1))$ представляє щільність багатовимірного нормального розподілу з параметрами $\mu_j(t\text{-}1)$ та $\Sigma_i(t\text{-}1)$.

На другому кроці алгоритму відбувається оновлення параметрів моделі з урахуванням обчислених ступенів належності та механізму забування. Для апріорних ймовірностей кластерів це реалізується за формулою:

$$\pi_{\mathrm{j}}(t) = \left(1 - \alpha(t)\right) \cdot \pi_{\mathrm{j}}(t-1) + \alpha(t) \cdot \mu_{\mathrm{j}}(x_{\mathrm{t}}), \quad (2.15)$$

де $\alpha(t) \in (0, 1)$ – коефіцієнт навчання, що може залежати від часу.

Ця формула реалізує експоненційне забування, за яким вплив спостережень зменшується з їх віком відповідно до параметра $(1\text{-}\alpha(t))$ [10].

Вектори середніх значень кластерів оновлюються за формулою:

$$\mu_{\mathrm{j}}(t) = \mu_{\mathrm{j}}(t-1) + \beta(t) \cdot \mu_{\mathrm{j}}(x_{\mathrm{t}}) \cdot \left(x_{\mathrm{t}} - \mu_{\mathrm{j}}(t-1)\right), \quad (2.16)$$

де $\beta(t) \in (0, 1)$ – коефіцієнт навчання для параметрів розподілу.

Ця формула представляє стохастичну апроксимацію рішення задачі максимізації правдоподібності та забезпечує поступову адаптацію центрів кластерів до нових спостережень з урахуванням їх нечіткої належності.

Алгоритм також містить механізми адаптивного визначення кількості кластерів, що дозволяє створювати нові кластери та об'єднувати існуючі залежно від еволюції структури даних у потоці. Створення нового кластера відбувається, коли поточне спостереження x_t має низьку правдоподібність відносно всіх існуючих кластерів. Об'єднання кластерів відбувається, якщо відстань між їх центрами стає меншою за порогове значення та їх форми є подібними.

Другий крок алгоритму – оновлення параметрів моделі з використанням обчислених ступенів належності та механізму забування. Для апріорних ймовірностей кластерів це виконується за формулою: $\pi_j(t) = (1-\alpha(t))\cdot\pi_j(t\text{-}1) + \alpha(t)\cdot\mu_j(x_t)$, де $\alpha(t) \in (0, 1)$ – коефіцієнт навчання, що може залежати від часу. Ця формула реалізує експоненційне забування, де вплив спостережень зменшується з їх віком відповідно до параметра $(1\text{-}\alpha(t))$ [10].

Оновлення векторів середніх значень кластерів виконується за формулою:

$$\mu_{\mathrm{j}}(t) = \mu_{\mathrm{j}}(t-1) + \beta(t)\cdot\mu_{\mathrm{j}}(x_{\mathrm{t}})\cdot\Big(x_{\mathrm{t}} - \mu_{\mathrm{j}}(t-1)\Big), \qquad (2.17)$$

де $\beta(t) \in (0, 1)$ – коефіцієнт навчання для параметрів розподілу.

Ця формула є стохастичною апроксимацією рішення задачі максимізації правдоподібності та забезпечує поступову адаптацію центрів кластерів до нових спостережень з урахуванням їх нечіткої належності.

Таблиця 2.3 – Параметри алгоритму правдоподібної нечіткої кластеризації для онлайн обробки

Параметр	**Позначення**	**Опис**	**Діапазон значень**	**Рекомен-довані значення**
1	**2**	**3**	**4**	**5**
Початкова кількість кластерів	c	Визначає кількість кластерів, з якої починає алгоритм	$c \geq 2$	2–5
Коефіцієнт навчання для апріорних ймовірнос-тей	$\alpha(t)$	Контролює швидкість забування старих ймовірностей та адаптації до нових	(0,1)	0,01–0,1
Коефіцієнт навчання для параметрів розподілу	$\beta(t)$	Контролює швидкість адаптації параметрів кластерів	(0,1)	0,001–0,01
Параметр регуляриза-ції для коваріацій-них матриць	γ	Запобігає виродженню коваріаційних матриць	> 0	0,01–0,1
Поріг для виявлення викидів	τ	Визначає граничну ймовірність для класифікації спостереження як викиду	(0,1)	0,001–0,01
Поріг для створення нового кластера	ρ	Мінімальна відстань від існуючих центрів для створення нового кластера	> 0	2–3 стандарт-них відхилення
Поріг для злиття кластерів	σ	Максимальна відстань між центрами для злиття кластерів	> 0	1–2 стандарт-них відхилення

Продовження таблиці 2.3

1	2	3	4	5
Розмір буфера пам’яті	*B*	Кількість останніх спостережень, що зберігаються для повторної кластеризації	≥ 0	100–1000

Оновлення коваріаційних матриць кластерів є найбільш обчислювально витратною операцією в алгоритмі та виконується за формулою:

$$\Sigma_{\mathrm{j}}(t) = \Sigma_{\mathrm{j}}(t-1) + \beta(t) \cdot \mu_{\mathrm{j}}(x_{\mathrm{t}}) \cdot \left(\left(x_{\mathrm{t}} - \mu_{\mathrm{j}}(t) \right) \cdot \left(x_{\mathrm{t}} - \mu_{\mathrm{j}}(t) \right)^{\mathrm{T}} - \Sigma_{\mathrm{j}}(t-1) \right), \quad (2.18)$$

Для забезпечення числової стабільності та запобігання виродженню матриць може застосовуватися регуляризація:

$$\Sigma_{\mathrm{j}}(t) = (1-\gamma) \cdot \Sigma_{\mathrm{j}}(t) + \gamma \cdot I \cdot \frac{tr\left(\Sigma_{\mathrm{j}}(t)\right)}{d}, \quad (2.19)$$

де $\gamma \in (0, 1)$ – параметр регуляризації;

I – одинична матриця;

$tr(\cdot)$ – слід матриці;

d – розмірність даних.

Механізм адаптивного визначення кількості кластерів становить основну складову алгоритму, дозволяючи формувати нові кластери та поєднувати наявні відповідно до змін у структурі потокових даних. Формування нового кластера здійснюється у випадку, коли поточне спостереження x_t демонструє недостатню правдоподібність стосовно всіх існуючих кластерів: $P(x_t|\theta(t\text{-}1)) = \sum_{j=1}^{c} \pi_j(t\text{-}1) \cdot f_j(x_t|\theta_j(t\text{-}1)) < \tau$, де τ представляє пороговий рівень. Ініціалізація нового кластера проводиться із центром у

точці x_t, коваріаційною матрицею, що пропорційна середній коваріації наявних кластерів, та незначною апріорною ймовірністю [21, 38].

Злиття кластерів відбувається за умови, що відстань між їхніми центрами стає меншою за встановлене порогове значення, а їхні форми виявляють подібність. З математичної точки зору, кластери i та j підлягають об'єднанню, якщо $\|\mu_i(t) - \mu_j(t)\|^2 < \sigma \cdot (tr(\Sigma_i(t))/d + tr(\Sigma_j(t))/d)$ та $|det(\Sigma_i(t))|^{(1/d)} / |det(\Sigma_j(t))|^{(1/d)} \in [1/r, r]$, де σ та r позначають порогові параметри. Параметри об'єднаного кластера обчислюються як зважені середні величини параметрів вихідних кластерів, з вагами, що відповідають їхнім апріорним ймовірностям.

Виявлення та опрацювання викидів становить невід'ємний аспект обробки потоків даних. Запропонований алгоритм відносить спостереження x_t до категорії викидів, якщо його правдоподібність щодо всіх кластерів є низькою, проте не настільки низькою, щоб створити новий кластер. Викиди не залучаються до процесу оновлення параметрів існуючих кластерів, що забезпечує стійкість алгоритму до шумів та аномалій у потоці даних [6].

Для підвищення ефективності адаптації до різких змін у структурі даних алгоритм може використовувати буфер пам'яті обмеженого розміру, що зберігає певну кількість останніх спостережень. При виявленні суттєвої зміни у характеристиках потоку (наприклад, значне зниження середньої правдоподібності спостережень) здійснюється повна перекластеризація даних з буфера для швидкого пристосування моделі до нової структури. Такий механізм виявляється особливо корисним при наявності раптових дрейфів концепцій у потоці даних.

Обчислювальна складність пропонованого алгоритму для обробки одного спостереження становить $O(c \cdot d^2)$, де c позначає кількість кластерів, а d – розмірність даних. Ця складність зумовлена операціями з коваріаційними матрицями розміру $d \times d$ для кожного з c кластерів. Для високорозмірних даних (значні величини d) можуть застосовуватися методи зниження

розмірності або спрощені моделі коваріаційних матриць, такі як діагональні матриці або матриці з факторною структурою.

Паралельна версія алгоритму може бути реалізована для розподіленої обробки потоків даних від численних джерел. При цьому для кожного джерела підтримується локальна модель, параметри якої періодично синхронізуються з глобальною моделлю. Синхронізація може виконуватися через зважене усереднення параметрів локальних моделей, де ваги визначаються відносною значущістю джерел або обсягом генерованих ними даних.

Для підвищення ефективності адаптації до різких змін у структурі даних алгоритм може використовувати буфер пам’яті обмеженого розміру, що зберігає B останніх спостережень. При виявленні суттєвої зміни у характеристиках потоку (наприклад, значне зниження середньої правдоподібності спостережень) виконується повна перекластеризація даних з буфера для швидкого пристосування моделі до нової структури. Цей механізм особливо корисний при наявності раптових дрейфів концепцій у потоці даних.

Обчислювальна складність запропонованого алгоритму для обробки одного спостереження становить $O(c \cdot d^2)$, де c – кількість кластерів, d – розмірність даних. Ця складність зумовлена операціями з коваріаційними матрицями розміру $d \times d$ для кожного з c кластерів. Для високорозмірних даних (великі значення d) можуть застосовуватися методи зниження розмірності або спрощені моделі коваріаційних матриць, такі як діагональні матриці або матриці з факторною структурою [24].

Паралельна версія алгоритму може бути реалізована для розподіленої обробки потоків даних від множинних джерел. При цьому для кожного джерела підтримується локальна модель, параметри якої періодично синхронізуються з глобальною моделлю. Синхронізація може виконуватися через зважене усереднення параметрів локальних моделей, де ваги

визначаються відносною важливістю джерел або обсягом даних, що вони генерують.

Алгоритм забезпечує інтерпретабельність результатів кластеризації через визначення нечітких ступенів належності спостережень до кластерів та параметрів розподілів, що описують кластери. Для гауссівської моделі це дозволяє візуалізувати кластери як еліпсоїди в просторі ознак, що відображають середні значення та коваріаційні структури даних у кластерах. Нечіткі межі між кластерами можуть бути візуалізовані через контури однакових рівнів функцій належності [36].

Експериментальна верифікація запропонованого алгоритму виконувалася на синтетичних та реальних потоках даних різної природи та складності. Синтетичні потоки генерувалися як послідовності спостережень з гауссівських розподілів з параметрами, що змінюються з часом, для моделювання різних сценаріїв дрейфу концепцій. Реальні потоки даних включали дані з сенсорів, мережевий трафік, фінансові часові ряди та текстові потоки з соціальних мереж. Результати експериментів підтверджують ефективність алгоритму з точки зору якості кластеризації, адаптивності до змін та обчислювальної ефективності.

Порівняння запропонованого алгоритму з існуючими методами нечіткої кластеризації потоків даних, такими як Online Fuzzy *C*-Means та Stream Possibilistic *C*-Means, показує його переваги у статистичній інтерпретації результатів, робастності до шуму та викидів, а також адаптивності до змін у структурі даних. Особливо значні переваги спостерігаються для потоків даних з комплексною структурою кластерів та наявністю дрейфу концепцій.

3 РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ РОЗРОБЛЕНОГО МЕТОДУ

3.1 Програмна реалізація методу правдоподібної нечіткої кластеризації

Програмна реалізація методу правдоподібної нечіткої кластеризації для онлайн обробки потоків даних поєднує статистичний підхід та нечітку логіку, враховуючи обмеження роботи в реальному часі. Розроблене програмне забезпечення написане мовою Python із застосуванням декількох спеціалізованих бібліотек: NumPy для векторних та матричних обчислень, SciPy для статистичних функцій, Pandas для маніпуляцій з даними та Matplotlib для створення візуалізацій. Архітектура програми побудована за модульним принципом, де кожен модуль відповідає за певний функціональний аспект алгоритму – від обробки вхідних даних до візуалізації результатів [3].

Модуль обробки вхідних даних забезпечує зчитування інформації з різноманітних джерел, таких як файли, мережеві джерела та потокові API. Він реалізує універсальний інтерфейс, що дозволяє працювати з різними джерелами даних через єдиний формат представлення. Кожне спостереження представляється вектором ознак у багатовимірному просторі та доповнюється часовою міткою надходження. Цей модуль також виконує попередню обробку даних – нормалізацію, видалення пропущених значень та фільтрацію аномалій, що сприяє покращенню результатів наступної кластеризації.

Центральну роль у програмній реалізації відіграє клас ProbabilisticFuzzyClustering, який містить основну логіку алгоритму та зберігає поточний стан моделі кластеризації. Цей клас включає структури для зберігання параметрів кластерів (вектори середніх значень, коваріаційні матриці, апріорні ймовірності) та допоміжні компоненти для накопичення статистик і оцінювання якості кластеризації. Клас реалізує методи для

ініціалізації моделі, обробки нових даних, розрахунку функцій правдоподібності та ступенів належності, оновлення параметрів моделі та адаптивного визначення кількості кластерів [39].

Функції правдоподібності реалізовані для різних типів розподілів, що надає гнучкість вибору найбільш відповідної моделі для конкретного завдання. Базова реалізація включає гауссівську функцію правдоподібності, що описується математичною формулою для багатовимірного нормального розподілу. Для підвищення обчислювальної ефективності програма розраховує логарифм функції правдоподібності безпосередньо, уникаючи обчислення експоненти, що зменшує ризик числового переповнення при роботі з багатовимірними даними.

Механізм нечіткої кластеризації реалізується через обчислення ступенів належності спостережень до кластерів на основі апостеріорних ймовірностей. Інкрементальне оновлення параметрів моделі відбувається за допомогою рекурентних формул, що дозволяють адаптувати модель без необхідності зберігати та повторно обробляти всі історичні дані. Така реалізація забезпечує ефективну роботу алгоритму в умовах обмежених обчислювальних ресурсів та пам’яті, що типово для систем онлайн-обробки потоків даних.

Функція правдоподібності реалізована для різних типів розподілів, що дозволяє користувачу вибирати найбільш підходящу модель для конкретної задачі. Базова реалізація включає гауссівську функцію правдоподібності, що описується формулою:

$$f(x|\mu,\Sigma) = (2\pi)^{\frac{-d}{2}}|\Sigma|^{\frac{-1}{2}}exp\big(-0{,}5(x-\mu)^T\Sigma^{-1}(x-\mu)\big), \qquad (3.1)$$

де x – вектор спостереження;

μ – вектор середніх значень;

Σ – коваріаційна матриця;

d – розмірність простору ознак.

Для підвищення обчислювальної ефективності логарифм функції правдоподібності обчислюється напряму, без обчислення експоненти, що зменшує ризик числового переповнення при роботі з багатовимірними даними.

Таблиця 3.1 – Структура програмної реалізації методу правдоподібної нечіткої кластеризації

Модуль	**Функціональність**	**Основні класи та методи**	**Використані бібліотеки**
data_stream	Взаємодія із джерелами даних	StreamReader, DataPreprocessor	Pandas, NumPy
probabilistic_model	Обчислення функцій правдоподібності	GaussianLikelihood, StudentLikelihood	NumPy, SciPy
fuzzy_clustering	Логіка нечіткої кластеризації	MembershipCalculator, FCMUpdater	NumPy
online_learning	Інкрементальне навчання	OnlineEM, AdaptiveRateCalculator	NumPy
cluster_adaptation	Адаптація кількості кластерів	ClusterMerger, ClusterSplitter	NumPy, SciPy
outlier_detection	Виявлення та обробка викидів	OutlierDetector	NumPy, SciPy
visualization	Візуалізація результатів	ClusterVisualizer, StreamVisualizer	Matplotlib, Seaborn
evaluation	Оцінка якості кластеризації	ValidationIndexCalculator	NumPy, SciPy

Механізм нечіткої кластеризації реалізований через обчислення ступенів належності спостережень до кластерів на основі апостеріорних ймовірностей.

Інкрементальне оновлення параметрів моделі реалізоване через рекурентні формули, що дозволяють адаптувати модель без необхідності

зберігання та повторної обробки всіх історичних даних. Для векторів середніх значень кластерів використовується формула:

$$\mu_{j(t)} = \mu_{j(t-1)} + \beta(t)\mu_{j(x_t)}(x_t - \mu_{j(t-1)}), \quad (3.2)$$

де $\beta(t)$ –коефіцієнт навчання, що може залежати від часу.

Аналогічні рекурентні формули використовуються для оновлення коваріаційних матриць та апріорних ймовірностей кластерів. Ці формули реалізують механізм експоненційного забування, де вплив спостережень зменшується з їх віком [21].

Система адаптивного визначення кількості кластерів функціонує завдяки двом основним процесам: формуванню нових та об'єднанню існуючих кластерів. Новий кластер створюється у випадках, коли система отримує спостереження з низькою правдоподібністю стосовно всіх наявних кластерів. Паралельно відбувається об'єднання кластерів за умови, що відстань між їхніми центрами не перевищує встановлений поріг, а форми демонструють схожість. Реалізація цих механізмів передбачає оптимальне використання обчислювальних ресурсів, що відповідає вимогам функціонування в режимі реального часу.

Програмна реалізація містить два методи забезпечення стійкості до викидів та шумів у даних. Перший метод зосереджується на виявленні та відфільтровуванні викидів перед їх застосуванням для оновлення параметрів моделі. Другий підхід впроваджує розподіли з «важчими хвостами», насамперед *t*-розподіл, замість традиційного нормального розподілу. Така заміна зменшує вплив віддалених спостережень на параметри кластерів.

Реалізація *t*-розподілу характеризується наявністю додаткового параметра степенів свободи, який дозволяє регулювати «важкість хвостів» розподілу. Суттєво, що цей параметр може динамічно налаштовуватись протягом процесу навчання, забезпечуючи адаптивність системи до характеристик вхідних даних.

Поєднання цих технік дає можливість створити надійну систему кластеризації, здатну ефективно працювати з даними, що містять шуми та викиди, зберігаючи при цьому точність визначення кластерної структури та обчислювальну ефективність, необхідну для обробки потоків даних у реальному часі [21].

Таблиця 3.2 – Параметри програмної реалізації методу правдоподібної нечіткої кластеризації

Параметр	**Тип**	**Опис**	**Значення за замовчуванням**
1	**2**	**3**	**4**
n_clusters	int	Початкова кількість кластерів	
distribution_type	str	Тип розподілу для моделювання кластерів	‘gaussian’
learning_rate	float	Базовий коефіцієнт навчання	0,01
forgetting_factor	float	Фактор забування для застарілих даних	0,99
regularization	float	Параметр регуляризації коваріаційних матриць	0,01
outlier_threshold	float	Порогове значення для виявлення викидів	0,001
creation_threshold	float	Порогове значення для створення нових кластерів	0,0001
merge_threshold	float	Порогове значення для об’єднання кластерів	2,0
buffer_size	int	Розмір буфера для адаптивної перекластеризації	100
max_clusters	int	Максимальна кількість кластерів	10
min_cluster_weight	float	Мінімальна вага кластера для збереження	0,01
covariance_type	str	Тип коваріаційної матриці	‘full’

Продовження таблиці 3.2

1	2	3	4
adaptive_rate	bool	Використання адаптивного коефіцієнта навчання	True
verbose	bool	Режим детального виведення інформації	False
random_state	int	Зерно для генератора випадкових чисел	None

Серед оптимізацій обчислень у програмній реалізації слід відзначити ефективну роботу з розрідженими матрицями, що зменшує вимоги до пам’яті при обробці високорозмірних даних; використання векторизованих операцій замість циклів для підвищення швидкодії; кешування проміжних результатів обчислень, що дозволяє уникнути повторних розрахунків; адаптивне обчислення коваріаційних матриць лише для спостережень з високими ступенями належності до відповідних кластерів [40].

Візуалізація результатів кластеризації реалізована через модуль, що дозволяє відображати еволюцію кластерів у часі, зміни їх параметрів, розподіл ступенів належності та інші характеристики процесу кластеризації [41]. Для багатовимірних даних реалізовані методи відображення проекцій на основні компоненти або обрані користувачем виміри. Анімація зміни структури кластерів у часі дозволяє наочно відслідковувати процес адаптації моделі до змін у потоці даних.

Програмна реалізація включає механізми серіалізації та десеріалізації моделі, що дозволяють зберігати поточний стан кластеризації та відновлювати його пізніше. Це є корисним для довготривалих експериментів, коли необхідно зберігати проміжні результати, а також для розподілених систем, де модель може переміщуватися між обчислювальними вузлами. Серіалізація реалізована з використанням стандартного модуля Python pickle, а також у форматі JSON для забезпечення інтероперабельності з іншими системами.

Для забезпечення масштабованості програмної реалізації при роботі з великими потоками даних використовуються механізми паралельних обчислень. Зокрема, реалізована можливість паралельного обчислення функцій правдоподібності для різних кластерів, що є найбільш обчислювально витратною операцією в алгоритмі. Для цього використовується бібліотека Python multiprocessing, що дозволяє розподіляти обчислення між кількома процесорними ядрами. Також реалізована можливість розподіленої обробки потоків даних з використанням Apache Kafka для передачі повідомлень та Apache Spark для паралельних обчислень [36].

Взаємодія з зовнішніми системами реалізована через API, що дозволяє інтегрувати розроблений метод у більші аналітичні системи. API включає методи для ініціалізації моделі, подачі нових спостережень, отримання поточних параметрів кластерів та ступенів належності, а також для управління процесом кластеризації. Реалізовані клієнтські бібліотеки для різних мов програмування (Python, Java, C++) спрощують інтеграцію у гетерогенні системи.

Документація програмної реалізації включає опис архітектури, API, прикладів використання та рекомендацій щодо налаштування параметрів для різних типів даних та задач. Документація створена з використанням системи Sphinx та доступна у форматах HTML та PDF. Також розроблені Jupyter ноутбуки з інтерактивними прикладами використання методу для різних наборів даних, що спрощує освоєння програмного забезпечення новими користувачами.

3.2 Експериментальна верифікація методу на тестових та реальних наборах даних

Експериментальна верифікація розробленого методу правдоподібної нечіткої кластеризації потоків даних проводилася на різноманітних наборах

даних, включаючи синтетичні тестові набори та реальні потоки даних. Синтетичні набори даних генерувалися з контрольованими характеристиками для оцінки здатності методу виявляти кластери різної форми, щільності, розміру, а також адаптуватися до змін у структурі даних. Реальні набори даних включали дані з різних доменів: мережевий трафік, фінансові часові ряди, сенсорні дані, текстові потоки з соціальних мереж. Загалом, для експериментів було використано 5 синтетичних та 8 реальних наборів даних різної природи та складності [3].

Для експериментальної верифікації методу було розроблено п’ять різних синтетичних наборів даних, сформованих за допомогою багатовимірних гауссівських розподілів із заданими параметрами. Набір S1 складався з п’яти статичних кластерів із чітко окресленими межами, що дозволило перевірити базові функціональні можливості методу щодо виявлення структури даних. Для дослідження ефективності нечіткої кластеризації був створений набір S2, що містив кластери з областями перетину між ними.

Набір S3 характеризувався наявністю кластерів складних геометричних форм (еліптичні та видовжені), за допомогою яких оцінювалась гнучкість розробленої моделі. Для тестування здатності системи до адаптації був розроблений набір S4, в якому параметри кластерів зазнавали поступових змін. Здатність методу пристосовуватися до різких трансформацій у структурі даних та сильних дрейфів концепцій перевірялася з використанням набору S5, що містив різкі структурні зміни в даних.

Оцінювання якості кластеризації на синтетичних наборах здійснювалося з застосуванням двох типів індексів валідності: зовнішніх та внутрішніх. Серед зовнішніх показників використовувались Adjusted Rand Index (ARI) та Normalized Mutual Information (NMI), розрахунок яких базувався на зіставленні результатів кластеризації з достовірними мітками кластерів.

Внутрішні індекси, включаючи нечіткий індекс Xie-Beni та нечіткий індекс ентропії, застосовувались для оцінки результатів кластеризації без залучення зовнішньої інформації. При роботі з потоками даних ці показники обчислювались у ковзному вікні, що забезпечило можливість моніторингу динаміки якості кластеризації в часі.

Такий комплексний підхід до формування тестових наборів даних та вибору метрик оцінювання дозволив всебічно проаналізувати функціональність розробленого методу в різних умовах, включаючи статичні структури, перетинні кластери, складні геометричні форми та динамічні зміни в даних. Результати експериментів надали повне уявлення про ефективність запропонованого методу та його межі застосування.

Таблиця 3.3 – Результати експериментів на синтетичних наборах даних

Набір даних	**Розмірність**	**Кількість кластерів**	**ARI**	**NMI**	**Xie-Beni**	**Ентропія**	**Час обробки(мс/спостереження)**
S1 (статичні кластери)	2	5	0,95	0,92	0,12	0,18	0,42
S2 (перетинні кластери)	3	4	0,82	0,79	0,25	0,35	0,58
S3 (складна форма)	5	6	0,78	0,76	0,31	0,40	0,87
S4 (поступова зміна)	4	3–7	0,85	0,81	0,28	0,33	0,74
S5 (різка зміна)	3	2–8	0,72	0,68	0,45	0,52	0,65

Експерименти на синтетичних наборах даних продемонстрували високу точність кластеризації для статичних кластерів (ARI = 0,95, NMI = 0,92) та задовільну точність для складніших сценаріїв, таких як кластери з перетинами (ARI = 0,82) та складної форми (ARI = 0,78). Метод показав

гарну адаптивність до поступових змін структури даних, зберігаючи високі показники якості кластеризації (ARI = 0,85). При різких змінах структури даних спостерігалося тимчасове зниження якості кластеризації, але метод швидко адаптувався до нової структури, відновлюючи прийнятний рівень точності.

Одним з ключових аспектів експериментальної верифікації була оцінка здатності методу адаптувати кількість кластерів відповідно до змін у структурі даних. Для наборів S4 та S5, де кількість кластерів змінювалася з часом, метод успішно відслідковував ці зміни, створюючи нові кластери при появі відокремлених груп спостережень та об’єднуючи кластери, що зблизилися. Середня похибка у визначенні кількості кластерів становила 0,8 для набору S4 та 1,2 для набору S5, що є прийнятним результатом для складних сценаріїв з дрейфом концепцій [10].

Для оцінки стійкості методу до шуму та викидів синтетичні набори даних були доповнені випадковими викидами з рівномірного розподілу в просторі ознак. Відсоток викидів варіювався від 5% до 20% загального обсягу даних. Експерименти показали, що метод зберігає прийнятний рівень точності кластеризації навіть при високому рівні шуму (20% викидів), з середнім зниженням ARI на 0,15 порівняно з наборами без шуму. Використання *t*-розподілу замість нормального підвищувало стійкість до викидів, зменшуючи середнє зниження ARI до 0,08.

Експериментальна верифікація методу правдоподібної нечіткої кластеризації проводилась на восьми реальних наборах даних, що представляли різні домени. Для цілісної оцінки практичної застосовності методу була обрана різноманітна структура тестових наборів.

Набір R1 складався з даних мережевого трафіку, що містили мітки атак, що дозволило перевірити здатність методу виявляти аномалії. У наборі R2 використовувались фінансові часові ряди з позначеними періодами різних ринкових режимів. Набір R3 представляв собою сенсорні дані зібрані системою моніторингу промислового обладнання, де також були присутні

мітки різних режимів роботи. Набір R4 формувався з потоків повідомлень із соціальних мереж, розмічених за тематикою. Додатково використовувались набори R5–R8, які містили дані з різних доменів без міток, що дозволило оцінити внутрішні характеристики процесу кластеризації [6].

Методика оцінки результатів розділилась на два підходи залежно від характеру даних. Для наборів з відомими мітками (R1–R4) застосовувались зовнішні індекси валідності, що дозволило виміряти відповідність знайдених алгоритмом кластерів реальним групам даних. При роботі з наборами без міток (R5–R8) оцінювання базувалось на внутрішніх індексах валідності та експертному аналізі виявлених кластерів спеціалістами у відповідних предметних областях.

Окрім точності кластеризації, проводилось вимірювання обчислювальної ефективності розробленого методу. Фіксувались показники часу обробки окремого спостереження та обсяг використаної пам’яті – ці параметри визначають можливість застосування методу для онлайн обробки потоків даних у реальних умовах. Такий комплексний підхід до тестування дозволив отримати повну картину продуктивності методу в різноманітних сценаріях використання та оцінити його практичну застосовність для різних типів задач кластеризації потоків даних.

Таблиця 3.4 – Результати експериментів на реальних наборах даних

Набір даних	**Домен**	**Розмір-ність**	**Виявлено кластерів**	**ARI/ NMI**	**Xie-Beni**	**Час обробки (мс/спосте-реження)**
1	**2**	**3**	**4**	**5**	**6**	**7**
R1	Мережевий трафік	12	7	0,75/ 0,72	0,31	1,25
R2	Фінансові часові ряди	8	4	0,68/ 0,65	0,38	0,92
R3	Сенсорні дані	18	5	0,81/ 0,78	0,22	1,84

Продовження таблиці 3.4

1	2	3	4	5	6	7
R4	Соціальні мережі	100	12	0,62/ 0,58	0,45	3,56
R5	Медичні дані	24	6	-/-	0,29	2,12
R6	Транспортні потоки	15	8	-/-	0,33	1,67
R7	Споживаць-ка поведінка	32	9	-/-	0,37	2,48
R8	Екологіч-ний моніторинг	10	4	-/-	0,26	1,05

Експерименти на реальних наборах даних показали дещо нижчу точність кластеризації порівняно з синтетичними наборами, що пояснюється складнішою структурою реальних даних та наявністю шуму. Найвища точність спостерігалася для набору сенсорних даних R3 (ARI = 0,81), найнижча – для набору даних соціальних мереж R4 (ARI = 0,62), що пояснюється високою розмірністю та складною семантичною структурою текстових даних. Загалом, метод продемонстрував здатність виявляти змістовні кластери у різних доменах, що було підтверджено як кількісними метриками, так і експертною оцінкою [24].

Для оцінки здатності методу адаптуватися до змін у реальних потоках даних аналізувалася динаміка внутрішніх індексів валідності та параметрів кластерів протягом часу. Зокрема, для фінансових часових рядів (R2) спостерігалася успішна адаптація методу до змін ринкових режимів, з коректним виявленням періодів високої та низької волатильності. Для даних соціальних мереж (R4) метод відслідковував зміни популярних тем з часом, виявляючи нові тематичні кластери при виникненні трендових тем.

Особливу увагу в експериментах було приділено оцінці обчислювальної ефективності методу, що є критичним аспектом для онлайн

обробки потоків даних. Вимірювався час обробки одного спостереження на стандартному обладнанні (процесор Intel Core i7, 16 ГБ оперативної пам’яті). Час обробки варіювався від 0,92 мс для набору R2 (8-вимірні дані) до 3,56 мс для набору R4 (100-вимірні дані), що дозволяє обробляти від 280 до 1087 спостережень за секунду. Такі показники є достатніми для багатьох практичних застосувань, включаючи аналіз мережевого трафіку та сенсорних даних у реальному часі.

Використання пам’яті методом залежало від розмірності даних, кількості кластерів та розміру буфера для адаптивної перекластеризації. Для набору R3 (18-вимірні дані, 5 кластерів, буфер розміром 100) використання пам’яті становило близько 5 МБ, що є прийнятним для онлайн систем. Для високорозмірних даних R4 (100 вимірів, 12 кластерів) використання пам’яті зростало до 25 МБ. Зменшення розміру буфера або використання спрощених моделей коваріаційних матриць (діагональних замість повних) дозволяло знизити використання пам’яті за рахунок незначного зниження точності кластеризації [3].

Серія експериментів для оцінки масштабованості розробленої методики включала тестування з варіативною інтенсивністю потоків даних і різною кількістю паралельних обчислювальних потоків. Результати засвідчили, що паралельна реалізація методу демонструє масштабованість, близьку до лінійної, при нарощуванні кількості процесорних ядер до восьми. Подальше збільшення кількості ядер супроводжувалося поступовим сповільненням зростання продуктивності, що пояснюється додатковими витратами обчислювальних ресурсів на забезпечення синхронізації між ядрами.

Тестування розподіленої версії методу на кластері з п’яти обчислювальних вузлів показало рівномірне зростання загальної пропускної здатності системи. Така конфігурація дозволила досягти швидкості обробки до 15000 спостережень за секунду при роботі з 8-вимірними наборами даних.

Порівняльний аналіз створеного методу з існуючими підходами до нечіткої кластеризації потоків даних (зокрема, з алгоритмами Online Fuzzy *C*-Means та Stream Possibilistic *C*-Means) проводився в уніфікованих умовах. Застосовувались ідентичні набори даних та єдині метрики оцінки якості кластеризації. Експериментальні дані засвідчили перевагу розробленого методу за показником точності кластеризації – спостерігалося середнє підвищення індексу ARI на 0,12 порівняно з OFCM та на 0,08 порівняно з SPCM.

Додатковими перевагами запропонованого методу виявились кращі характеристики адаптивності до структурних змін у потоках даних та підвищена стійкість до негативного впливу шумів і викидів. Це підтверджує доцільність застосування статистичного підходу на основі функції правдоподібності, що дозволяє ефективніше працювати з неоднорідними потоками даних, які містять шумові компоненти.

Результати проведених порівняльних експериментів свідчать, що розроблений метод забезпечує збалансоване поєднання високої точності кластеризації з прийнятними показниками обчислювальної ефективності, що робить його придатним для застосування в різноманітних задачах аналізу потоків даних.

3.3 Аналіз ефективності та обмежень розробленого методу

Аналіз ефективності розробленого методу правдоподібної нечіткої кластеризації потоків даних в онлайн режимі проводився за кількома ключовими аспектами: точність кластеризації, адаптивність до змін у структурі даних, стійкість до шуму та викидів, обчислювальна ефективність та масштабованість. Порівняльний аналіз з існуючими методами дозволив виявити сильні сторони та обмеження розробленого підходу. Метод демонструє високу точність кластеризації для статичних наборів даних (ARI = 0,95 для синтетичного набору S1) та задовільну точність для динамічних

потоків з дрейфом концепцій (ARI = 0,72–0,85 для наборів S4–S5). Ці показники перевищують аналогічні для таких методів, як Online Fuzzy *C*-Means (OFCM) та Stream Possibilistic *C*-Means (SPCM), що підтверджує ефективність статистичного підходу на основі функції правдоподібності.

Адаптивність до змін у структурі даних є однією з сильних сторін розробленого методу. Механізми створення та об'єднання кластерів, а також експоненційне забування застарілих даних забезпечують ефективну адаптацію до поступових та різких змін. Час адаптації після різких змін у структурі даних становить у середньому 35 спостережень для синтетичних наборів та 50–120 спостережень для реальних потоків даних, що є кращим показником порівняно з аналогічними методами. Експерименти на реальних наборах даних з виразними змінами структури, таких як фінансові часові ряди (R2) та потоки соціальних мереж (R4), підтверджують здатність методу відслідковувати еволюцію кластерів з часом [3].

Стійкість до шуму та викидів забезпечується використанням статистичних розподілів з «важкими хвостами» та механізмами виявлення аномалій. Експерименти з додаванням контрольованого рівня шуму (від 5% до 20% викидів) показали, що метод зберігає прийнятний рівень точності кластеризації, з середнім зниженням ARI на 0,15 для гауссівської моделі та лише на 0,08 для моделі на основі *t*-розподілу. Це значно краще, ніж для методу OFCM, де зниження ARI при 20% викидів становило 0,28. Здатність ідентифікувати та ізолювати викиди, не допускаючи їх впливу на параметри моделі, є перевагою розробленого методу в умовах зашумлених даних.

Таблиця 3.5 – Порівняння методу правдоподібної нечіткої кластеризації з існуючими методами

Метод	**Точність (ARI)**	**Час адаптації (спост.)**	**Стійкість до шуму (зниження ARI при 20% викидів)**	**Обчислювальна складність**	**Пам'ять (МБ для R3)**
Розроблений метод (гауссівський)	0,81	35	0,15	$O(c \cdot d^2)$	5,2
Розроблений метод (t-розподіл)	0,79	38	0,08	$O(c \cdot d^2)$	5,5
Online Fuzzy *C*-Means	0,69	62	0,28	$O(c \cdot d)$	3,8
Stream Possibilistic *C*-Means	0,73	48	0,19	$O(c \cdot d)$	4,1
Growing Neural Gas	0,75	45	0,22	$O(c^2 \cdot d)$	6,3
DenStream	0,77	52	0,12	$O(n \cdot d)$	8,7

Обчислювальна складність методу становить $O(c \cdot d^2)$, де c – кількість кластерів, d – розмірність даних. Це дещо вище, ніж для методів з обчислювальною складністю $O(c \cdot d)$, таких як OFCM та SPCM, але нижче, ніж для методів на основі щільності, таких як DenStream з складністю $O(n \cdot d)$, де n – кількість точок у локальній околиці. На практиці час обробки одного спостереження варіюється від 0,92 мс до 3,56 мс залежно від розмірності даних, що дозволяє обробляти від 280 до 1087 спостережень за секунду на стандартному обладнанні. Ці показники є достатніми для багатьох практичних застосувань, хоча можуть бути обмеженням для надзвичайно інтенсивних потоків даних.

Аналізуючи ефективність методу щодо використання пам’яті, слід зазначити, що для стандартних наборів даних з розмірністю 10–20 та кількістю кластерів 5–10 споживання пам’яті знаходиться в межах 5–10 МБ. Такі показники цілком відповідають потребам онлайн систем та не створюють надмірного навантаження. Проте ситуація змінюється при роботі з високорозмірними даними, особливо коли кількість вимірів перевищує 100, а кількість кластерів зростає. У таких випадках споживання пам’яті може досягати десятків мегабайтів, що становить певну проблему для систем з обмеженим обсягом оперативної пам’яті.

Одним із шляхів зменшення навантаження на пам’ять є застосування спрощених моделей коваріаційних матриць. Діагональні та факторні матриці потребують значно менше пам’яті порівняно з повними коваріаційними матрицями [42]. Хоча таке спрощення призводить до певного зниження точності кластеризації, компроміс між економією пам’яті та незначною втратою точності може бути прийнятним для багатьох практичних застосувань.

Щодо масштабованості, метод демонструє задовільні характеристики як для паралельної обробки на одному вузлі, так і для розподіленої обробки на кластері. Паралельна версія показує майже лінійне зростання продуктивності при збільшенні кількості процесорних ядер до 8. Після цього порогового значення спостерігається поступове сповільнення зростання продуктивності.

Розподілена версія методу також демонструє хороші характеристики масштабованості. При збільшенні кількості вузлів пропускна здатність зростає пропорційно, що дозволяє обробляти тисячі спостережень за секунду. Однак при роботі з дуже великими кластерами виникає проблема синхронізації параметрів моделі між вузлами. Ця необхідність синхронізації створює додаткові накладні витрати, які негативно впливають на загальну ефективність масштабування [36].

Таблиця 3.6 – Обмеження методу правдоподібної нечіткої кластеризації та можливі шляхи їх подолання

Обмеження	**Опис**	**Шляхи подолання**	**Потенційний вплив на ефективність**
1	**2**	**3**	**4**
Висока обчислювальна складність для високорозмірних даних	Складність $O(c \cdot d^2)$ стає проблематичною при $d > 100$	Зниження розмірності, спрощені моделі коваріації	Можливе зниження точності на 5–10%
Чутливість до вибору початкових параметрів	Результати кластеризації можуть залежати від ініціалізації	Множинні запуски з різними ініціалізаціями, метод «розщеплення-злиття»	Збільшення обчислювальних витрат на 20–30%
Складність визначення оптимальної кількості кластерів	Автоматичне визначення може бути неточним для складних даних	Байєсівський підхід з апріорним розподілом для кількості кластерів	Збільшення обчислювальних витрат на 15–25%
Обмежена ефективність для потоків з нестаціонарним шумом	Характеристики шуму можуть змінюватися з часом	Адаптивні моделі для виявлення та обробки викидів	Незначне збільшення обчислювальних витрат
Затримка адаптації при різких змінах	Потрібен час для адаптації після сильних дрейфів концепцій	Буфер для перекластеризації, механізми виявлення змін	Збільшення використання пам’яті
Обмежена інтерпретабельність нечітких кластерів	Складно інтерпретувати кластери з перетинами	Візуалізація, опис кластерів через прототипи	Додаткова обчислювальна складність

Продовження таблиці 3.6

1	2	3	4
Масштабова ність для надвисоких інтенсивнос-тей потоків	Обмеження при обробці мільйонів спостережень за секунду	Апроксимаційні методи, вибіркова обробка	Можливе зниження точності
Підтримка неструктуро-ваних даних	Обмежена ефективність для текстів, зображень без попередньої обробки	Інтеграція з методами глибокого навчання	Збільшення обчислюваль-ної складності

Серед обмежень розробленого методу потрібно відзначити залежність від вибору початкових параметрів. Результативність процесу кластеризації прямо залежить від стартової кількості кластерів, їх первинного розташування та додаткових параметрів моделі. Незважаючи на наявність механізмів, що адаптують кількість кластерів у процесі роботи, повністю нівелювати вплив початкових умов неможливо. Проведені експериментальні дослідження демонструють, що використання різних варіантів ініціалізації призводить до розбіжностей показника ARI в межах 0,05–0,10. Особливо ця тенденція проявляється при обробці складних даних, що характеризуються кластерами з перетинами та розмитими межами. Для подолання цього обмеження можна використовувати підхід з багаторазовими запусками при різних початкових параметрах з подальшим вибором оптимального результату на основі внутрішніх індексів валідності. Проте такий метод потребує додаткових обчислювальних ресурсів.

Не менш проблематичним залишається питання визначення оптимальної кількості кластерів, особливо коли йдеться про потоки даних зі змінною структурою. Розроблений метод передбачає механізми для створення нових кластерів та об’єднання існуючих, однак при роботі з надзвичайно комплексними даними точність автоматичного визначення їх кількості може бути недостатньою. Аналіз роботи алгоритму на синтетичних

даних із заздалегідь відомою кількістю кластерів показав середню похибку на рівні 0,8–1,2 кластера, що можна вважати прийнятним результатом. Водночас для реальних наборів даних без чітко вираженої структури визначення «правильної» кількості кластерів часто носить суб’єктивний характер і значною мірою залежить від конкретного застосування [24].

Питання щодо нестаціонарних шумових характеристик варто розглянути детальніше. Запропонований метод нечіткої кластеризації показує надійну стійкість до стаціонарного шуму, проте його результативність може погіршуватись при зміні характеристик шуму з плином часу. Коли розподіл аномальних значень трансформується, наприклад, з рівномірного стає концентрованим, це може спричинити помилкову ідентифікацію кластерів у шумових даних. Для подолання цих обмежень доцільно використовувати адаптивні моделі виявлення та обробки викидів, які враховують темпоральні зміни в шумових характеристиках, що дозволить підвищити робастність методу в подібних сценаріях.

Інше обмеження пов’язане із затримкою адаптації при раптових змінах у структурі даних. Це типова проблема для інкрементальних методів аналізу. Хоча розроблений підхід демонструє швидшу адаптацію порівняно з існуючими аналогами, все ж існує певний лаг (від 35 до 120 спостережень) до повного пристосування після значних дрейфів концепцій. У цей період точність кластеризації неминуче знижується.

Для мінімізації такої затримки можна застосувати спеціальний буфер для повторної кластеризації та впровадити механізми, які цілеспрямовано виявляють структурні зміни в даних. Однак впровадження таких рішень потребує додаткових обчислювальних потужностей та збільшує об’єм використовуваної пам’яті.

Спостереження показують, що різкі зміни в структурі даних становлять особливий виклик для методів онлайн-кластеризації. Адаптація параметрів кластерів до нових патернів вимагає накопичення достатньої кількості спостережень, що неминуче створює тимчасовий період зниженої точності

результатів. Цей компроміс між швидкістю адаптації та обчислювальною ефективністю залишається відкритим питанням у галузі аналізу потоків даних.

Перевагою запропонованого методу є оптимізований час адаптації, але навіть з цим покращенням період пристосування залишається помітним при суттєвих змінах у даних. Збалансування часу адаптації та ефективності використання ресурсів є одним із напрямків для подальшого розвитку методу [24].

Нечіткі кластери часто стикаються з обмеженнями в інтерпретабельності, особливо коли дані містять значні перетини між кластерами. Незважаючи на те, що нечітка кластеризація забезпечує гнучкіший опис структури даних, розуміння ступенів належності та визначення меж між кластерами становить серйозну проблему для користувачів, які не мають спеціалізованих знань у цій галузі. Для подолання цієї проблеми можна застосовувати методи візуалізації кластерів та описувати їх через характерні прототипи – найбільш репрезентативні спостереження. Такий підхід сприяє кращому розумінню структури кластерів, проте потребує залучення додаткових інструментів та методик [43]

При роботі з надзвичайно інтенсивними потоками даних (мільйони спостережень за секунду) питання масштабованості стає суттєвим обмеженням. Розподілена версія методу демонструє здатність обробляти до 15000 спостережень за секунду на кластері з 5 вузлів. Ця продуктивність задовольняє вимоги багатьох практичних застосувань, однак виявляється недостатньою для екстремальних сценаріїв – аналізу мережевого трафіку на магістральних каналах чи обробки даних з високочастотних фінансових торгів.

Для розв’язання проблеми масштабованості доцільно застосовувати апроксимаційні методи. Серед таких підходів виділяються стохастична мініпакетна обробка та вибіркова обробка, базована на оцінці значущості спостережень. Ці методи дозволяють суттєво підвищити пропускну здатність

системи, хоча й можуть призводити до певного зниження точності результатів.

Обидва описані обмеження – інтерпретабельність нечітких кластерів та масштабованість для надінтенсивних потоків – потребують компромісних рішень. У першому випадку компроміс полягає між складністю представлення даних та зрозумілістю для користувача, у другому – між швидкістю обробки та точністю аналізу. Вибір оптимального балансу залежить від конкретних вимог предметної області та наявних обчислювальних ресурсів.

Підтримка неструктурованих даних, таких як тексти, зображення, аудіо, без попередньої обробки та виділення ознак є обмеженою. Метод працює безпосередньо з векторами ознак у числовому просторі, тому для неструктурованих даних необхідні додаткові етапи перетворення даних у відповідний формат. Інтеграція з методами глибокого навчання для автоматичного виділення ознак може розширити застосовність методу до ширшого спектру даних, але збільшує обчислювальну складність та вимоги до навчальних даних.

Баланс між точністю та ефективністю є фундаментальним обмеженням, що вимагає компромісів при налаштуванні методу для конкретних застосувань. Підвищення точності кластеризації зазвичай вимагає більш складних моделей з більшою кількістю параметрів, що збільшує обчислювальні витрати та використання пам’яті. Навпаки, підвищення ефективності шляхом спрощення моделей може призводити до зниження точності. Вибір оптимального балансу залежить від конкретного застосування, вимог до точності та доступних обчислювальних ресурсів [24].

У підсумку, аналіз ефективності та обмежень розробленого методу правдоподібної нечіткої кластеризації потоків даних в онлайн режимі показує, що він забезпечує збалансовані показники точності, адаптивності, стійкості до шуму та обчислювальної ефективності для широкого спектру застосувань. Виявлені обмеження можуть бути адресовані через додаткові

механізми та модифікації, що адаптують метод до специфічних вимог конкретних задач. Подальший розвиток методу може бути спрямований на подолання цих обмежень та розширення його застосовності до більш складних сценаріїв та типів даних.

ВИСНОВКИ

У результаті проведеної роботи розроблено та математично обґрунтовано метод правдоподібної нечіткої кластеризації потоків даних в онлайн режимі, що забезпечує ефективний аналіз неперервних послідовностей спостережень в умовах обмежених обчислювальних ресурсів.

Проведений аналіз існуючих методів нечіткої кластеризації та їх адаптацій для роботи з потоками даних дозволив визначити основні напрямки вдосконалення підходів до кластеризації неперервних послідовностей. Виявлено, що поєднання принципів статистичної теорії оцінювання та нечіткої логіки створює потужний апарат для моделювання складних структур у даних з урахуванням невизначеності та статистичної природи інформації.

Розроблена математична модель методу правдоподібної нечіткої кластеризації потоків даних базується на інкрементальній версії алгоритму очікування-максимізації та використовує функцію правдоподібності для формування кластерів. Новизна запропонованого підходу полягає у статистично обґрунтованому механізмі оцінки параметрів кластерів, адаптивному механізмі забування застарілих даних, робастності до шуму та ефективних механізмах адаптації до змін структури даних у часі.

Програмна реалізація розробленого методу виконана на мові програмування Python з використанням бібліотек NumPy, SciPy, Pandas та Matplotlib. Архітектура програмного забезпечення структурована у вигляді модулів, що відповідають за різні аспекти алгоритму та забезпечують можливість інтеграції з зовнішніми системами через реалізоване API.

Експериментальна верифікація методу на синтетичних та реальних наборах даних різної природи та складності підтвердила його ефективність за критеріями точності кластеризації, адаптивності до змін та обчислювальної ефективності. Зокрема, метод продемонстрував високу точність кластеризації для статичних кластерів (ARI = 0,95) та задовільну точність для динамічних

потоків з дрейфом концепцій (ARI = 0,72–0,85). Час обробки одного спостереження варіюється від 0,92 мс до 3,56 мс залежно від розмірності даних, що дозволяє обробляти від 280 до 1087 спостережень за секунду на стандартному обладнанні.

Порівняльний аналіз з існуючими методами нечіткої кластеризації потоків даних показав переваги розробленого підходу у статистичній інтерпретації результатів, робастності до шуму та викидів, а також адаптивності до змін у структурі даних. Розроблений метод демонструє вищу точність кластеризації порівняно з Online Fuzzy *C*-Means (середнє підвищення ARI на 0,12) та Stream Possibilistic *C*-Means (середнє підвищення ARI на 0,08).

Виявлено обмеження методу, що включають чутливість до вибору початкових параметрів, складність визначення оптимальної кількості кластерів, затримку адаптації при різких змінах структури даних, обмежену інтерпретабельність нечітких кластерів та масштабованість для надзвичайно високих інтенсивностей потоків. Запропоновано шляхи подолання цих обмежень через додаткові механізми та модифікації.

Розроблений метод правдоподібної нечіткої кластеризації потоків даних може бути застосований у різних галузях, де виникає необхідність аналізу неперервних послідовностей спостережень у реальному часі, включаючи моніторинг мережевого трафіку, аналіз даних соціальних мереж, обробку сенсорних даних та аналіз фінансових транзакцій.

Результати роботи розширюють теоретичні основи інтелектуального аналізу даних та створюють підґрунтя для подальшого розвитку методів кластеризації та класифікації даних в умовах невизначеності та обмежених обчислювальних ресурсів. Перспективними напрямками подальших досліджень є інтеграція розробленого методу з глибоким навчанням, вдосконалення методів для обробки неструктурованих та мультимодальних даних, а також розробка більш ефективних механізмів адаптації до змін у структурі даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Колінько П. (2022). Розпізнавання сонячних панелей на супутникових зображеннях. https://ekmair.ukma.edu.ua/server/api/core/bitstreams/15bc3d66-39da-442a-bcf6-0597152d94b6/content

2. Авлякулов Т. Е. (2023). Аналіз методів кластеризації на основі щільності розподілу даних. https://openarchive.nure.ua/entities/publication/b34ada2f-f3ae-4eaa-9160-7dd66ee5e016

3. Сараєв О. В. (2021). Еволюційна нейро-фаззі система з online навчанням-самонавчанням. https://openarchive.nure.ua/server/api/core/bitstreams/36b96d63-5714-4ef3-87ce-079f2828d631/content

4. Ланде Д. В., Субач І. Ю., Бояринова Ю. Є. (2018). Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки. https://ela.kpi.ua/server/api/core/bitstreams/5eb4cb50-3ef0-44d1-b35f-a80eade1554b/content

5. Полібіна К. В. (2024). Сценарно-прецеденті моделі підтримки індивідуальної освітньої траєкторії підготовки фахівців з інженерії програмного забезпечення. https://eir.kntu.net.ua/jspui/handle/123456789/1625

6. Ковальов І. Д. (2018).Оцінка ризиків інформаційної безпеки з використанням алгоритму нечіткої кластеризації k-середніх. https://core.ac.uk/download/pdf/168411939.pdf

7. Солнцев С. О., Жигалкевич Ж. М. (2023). Моделювання управління економічними системами і процесами. Навчально-методичний комплекс дисципліни (перероблений і доповнений). https://ela.kpi.ua/items/13a7c8a6-d91f-4cb7-b393-966e23dfe406

8. Пеня О. Р.(2021). Алгоритмічно-програмний метод агрегації даних цифрових двійників.

https://ela.kpi.ua/server/api/core/bitstreams/64c8c1cc-f580-48a0-8fb4-255607ba9bed/content

9. Прокопенко О. Є. (2025). Розробка навчальної системи для збору персональних з використанням штучного інтелекту та соціальної інженерії : магістерська дис. ... https://elartu.tntu.edu.ua/handle/lib/48324

10. Шафроненко А. Ю. (2023). Адаптивні методи нечіткої кластеризації потоків даних з використанням еволюційного самонавчання. https://openarchive.nure.ua/entities/publication/f54b7471-049e-4dcd-a034-ee959220c9fa

11. Михайлишин Д. А. (2022). Система моніторингу безпеки кінцевих пристроїв : дис. ... http://dspace.wunu.edu.ua/bitstream/316497/46655/1/Михайлишин.pdf

12. Плаксицький В. А. (2024). Методи і засоби кластеризації територій за індикаторами загроз енергетичної безпеки. https://ela.kpi.ua/items/bfac6f02-b9f1-48bc-b249-8f85708e8356

13. Бояринцев О. О. (2023). Розробка алгоритмів діагностування на основі нечіткої кластеризації даних. http://eprints.library.odeku.edu.ua/id/eprint/12009/3/BoyarintsevO_B_2023.pdf

14. Єрьоменко В. М. (2021). Дослідження методу структурної класифікації зображень з використанням засобів нечіткої кластеризації даних. https://openarchive.nure.ua/entities/publication/fd03758f-610f-4d05-ba70-73042e8b3f28

15. Іванов С. І. (2021). Прогнозування результатів виборів за допомогою нейронних мереж. https://ela.kpi.ua/server/api/core/bitstreams/5c0ec2e7-10dc-4440-8b7d-63a2a4e38e84/content

16. Касумов А. І. (2025). Дослідження та реалізація методу розпізнавання голосових команд за допомогою нейромережі. https://openarchive.nure.ua/entities/publication/17141d09-2c70-4212-af81-58a22b827543

17. Личагіна С. М. (2025). Дослідження та моделювання нейрофаззі мережі Кохонена для кластеризації викривлених даних. https://openarchive.nure.ua/entities/publication/c459f21e-6837-48ce-a17b-3047eb0f711a

18. Рипіч Б. В. (2024). Метод управління проєктами з розробки програмного забезпечення на основі нечіткої логіки та нейронних мереж : дис. ... Тернопіль : ЗУНУ, http://dspace.wunu.edu.ua/bitstream/316497/53331/1/Рипіч.pdf

19. Кучеренко А. А., Лисенко О. І., Новіков В. І. (2022). Кластеризація в безпроводових сенсорних мережах з використанням нечіткої логіки. Збірник матеріалів Міжнародної науково-технічної конференції «ПЕРСПЕКТИВИ ТЕЛЕКОМУНІКАЦІЙ». С. 189-191.

20. Яременко Є. А. (2018). Система орієнтування на місцевості з використанням технології доповненої реальності. https://ela.kpi.ua/server/api/core/bitstreams/f1d3c2d6-e4d1-4257-b34c-8b4a34e74c64/content

21. Бодянський Є. В., Шафроненко А. Ю., Климова І. М. (2021). Онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй. Radio Electronics, Computer Science, Control. № 2. С. 65-70. URL: https://ric.zp.edu.ua/article/view/236116

22. Магніцький Є. Д. (2025).Дослідження та реалізація методу нечіткої сегментації зображень на основі теорії достовірності. https://openarchive.nure.ua/entities/publication/4ee0cc06-b92d-48db-a250-2f20f1d58a49

23. Антіпова К. О. (2020). Моделі та інформаційні технології аналізу і вибору складних лінійно-функціональних організаційних структур в умовах невизначеності. https://dspace.chmnu.edu.ua/jspui/handle/123456789/1252

24. Кобилін І. О. (2019). Нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних.

https://openarchive.nure.ua/server/api/core/bitstreams/4d3350e1-2c3b-44b8-89af-fa0092519d45/content

25. Кобилін І. О. (2019). Нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних. https://openarchive.nure.ua/server/api/core/bitstreams/4d3350e1-2c3b-44b8-89af-fa0092519d45/content

26. Харченко В. В. (2019). Розробка та дослідження адаптивного методу ймовірнісної нечіткої кластеризації даних. https://openarchive.nure.ua/entities/publication/5ef105b9-13f9-4736-adde-6ce8cd5ac4cc

27. Юрчук М. В. (2019). Система аналізу та категоризації текстових медичних даних з використанням SAS технологій. https://ela.kpi.ua/server/api/core/bitstreams/86c7091e-3cc3-430a-9adc-6f103508d3d2/content

28. Стеблянко Б. О. (2024). Метод нечіткої кластеризації коротких текстів. https://openarchive.nure.ua/entities/publication/3aa038e0-3ddb-4dc2-808c-999d071a1d8a

29. Стрела Т. С. (2018) Метод вибору головного вузла кластеру в безпроводових сенсорних мережах з використанням нечіткої логіки. Збірник наукових праць [Військового інституту телекомунікацій та інформатизації]. № 4. С. 113-124.

30. Фалько М. К. (2025). Дослідження та реалізація еволюційного методу кластеризації даних. https://openarchive.nure.ua/entities/publication/2139f303-e2ba-4969-98f3-62be0cd70d47

31. Ахрамович В. М. (n.d.) Методологічні основи захисту інформації в соціальних мережах. https://duikt.edu.ua/uploads/p_1539_46668124.pdf

32. Вискребенцева С. О. (2019). Розробка та дослідження методу кластеризації для прогнозування.

https://openarchive.nure.ua/entities/publication/35f808c8-6faa-42c1-a8d3-3ace8d247c9e

33. Романенко А. Р. (2021). Послідовне нечітке кластерування потоків даних на основі ансамблевого підходу. https://openarchive.nure.ua/entities/publication/6e9a58b5-83e0-4099-9597-47dc8d42fb5f

34. Голубничий О. Г. (2019). Аналіз особливостей реалізації ЕМ-алгоритму при кластеризації систем сигнальних конструкцій. Наукоємні технології. № 2. С. 246-253. http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21REF=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=Nt_2019_2_14

35. Гончар Я. А. (2024). Методи відновлення відсутніх даних на основі нейронних мереж : дис. ... Тернопіль : ЗУНУ, http://dspace.wunu.edu.ua/handle/316497/53319

36. Гороховатський В. О., Творошенко І. С. (2021). Методи інтелектуального аналізу та оброблення даних : навч. посібник. https://openarchive.nure.ua/entities/publication/97cb0923-d711-4d14-afca-56d3f092b562

37. Ваховська О. В. (2023) Прикладні аспекти комп'ютерної лінгвістики. http://library.megu.edu.ua:8180/jspui/handle/123456789/4810

38. Безверха Є. В. (2024). Дослідження методів відновлення пошкоджених даних. https://openarchive.nure.ua/entities/publication/93c05f8c-af04-45b3-aa46-b93fea845dc6

39. Ріпний В. В. (2023) Розробка методу видалення рукописних написів на фотографії. https://openarchive.nure.ua/entities/publication/080e4e8c-a9a5-4a88-9d72-305a29a112d0

40. Пригодій А. І. (2020). Дослідження та реалізація методу машинного навчання для класифікації великих обсягів даних.

https://openarchive.nure.ua/server/api/core/bitstreams/95bcc1ff-4d3c-4038-91bb-6837d4906cb9/content

41. Науменко В. О. (2021). Методи генерації палітри кольорів зображення за допомогою інтелектуальних систем. https://ela.kpi.ua/server/api/core/bitstreams/37ad56fd-16f8-4bc6-8d92-8b3952f8f7ad/content

42. Фалько М. К. (2023). Розробка та моделювання методу онлайн-кластеризації даних за умов перетинних кластерів. https://openarchive.nure.ua/server/api/core/bitstreams/7eb67784-a347-4f1c-a022-db35655908d4/content

43. Удовенко С. Г., Келембет Д. В., Тесленко О. В. (2020) Комбінований метод нечіткої кластеризації даних в системах технічної діагностики. Системи обробки інформації. № 1 (160). С. 7-17. https://d1wqtxts1xzle7.cloudfront.net/93319136/114-libre.pdf?1667143177=&response-content-disposition=inline%3B+filename%3D93319136.pdf&Expires=1744480104&Signature=cagYR4dn~yQ43CMGl6ejGh3pTQW4002X2xzpxkdpiPtsQC8nQ2QhhKppWq350T2Gu3YYIOhfswQYDe4ggRdR3GS2VnatgbeOeFbJU-FOrYFSdKhjGPhsU42WwkWNcYCIAeg-1Cu1R3bIVWNa0D1XisDClN5HwP2wB17okukxlNY7TEA2mi2zgBf0eyzeL7ubgKwlakKEcMVCytTNfd9OvovnEaaKAyfKHdM3qTsCcx3KWFxN6DFrSMXdo-Z6I1ZWSS9eFPHEDw1481ycYtA7UzlinwBYgLpzM14Uq3UpHOdgclkqsVHnB353AclGIK87NXAw53GTowZrh2lIiFQGyw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA